



**LEVEL #**

12

Review of Test Theory and Methods

AD A 096157

David J. Weiss

and

Mark L. Davison

DTIC  
ELECTIC  
S MAR 10 1981 D

A

RESEARCH REPORT 81-1  
JANUARY 1981

COMPUTERIZED ADAPTIVE TESTING LABORATORY  
PSYCHOMETRIC METHODS PROGRAM  
DEPARTMENT OF PSYCHOLOGY  
UNIVERSITY OF MINNESOTA  
MINNEAPOLIS, MN 55455

This review was supported by funds from the Army Research Institute, the Air Force Human Resources Laboratory, the Air Force Office of Scientific Research, and the Office of Naval Research, and monitored by the Office of Naval Research.

Approved for public release; distribution unlimited.  
Reproduction in whole or in part is permitted for any purpose of the United States Government.

DBG FILE COPY

81 3 09 018

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 81-1	2. GOVT ACCESSION NO. AD-A096157	3. RECIPIENT'S CATALOG NUMBER (14) RR-81-1
4. TITLE (and Subtitle) Review of Test Theory and Methods.		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) David J. Weiss and Mark L. Davison	15	8. CONTRACT OR GRANT NUMBER(s) N00014-79-C-0172
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Minnesota Minneapolis, Minnesota 55455	17	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.: 6115N Proj.: RR042-04 T.A.: RR042-04-01 W.U.: NR 150-433
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217	11	12. REPORT DATE Jan 1981
		13. NUMBER OF PAGES 61
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES This paper is an extended version of a review prepared for the Annual Review of Psychology, Volume 32, pages 629-658. Preparation of this paper was supported by funds to the first author from the Army Research Institute, the Air Force Human Resources Laboratory, the Air Force Office of Scientific Research, and the Office of Naval Research, and monitored by the Office of Naval Research; and by a National Academy of Education Spencer Fellowship to the second author.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) classical test theory    validity    criterion-referenced testing latent trait test theory    reliability    multitrait-multimethod matrices item response theory    test bias    aptitude testing order theory    item types    ability testing generalizability theory    response modes    achievement testing		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) The research literature on test theory and methods for the period 1975 through early 1980 is critically reviewed. Research on classical test theory has concentrated on relatively unimportant developments in reliability theory, with some new developments and applications of generalizability theory appearing during this period. The reliability of change or gain scores has received some attention from the classical test theory perspective, as have the applications of classical reliability concepts in experimental design and the anal-		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 65 IS OBSOLETE  
S/N 0102-LF-014-6601

Unclassified 4060-4

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ysis of experimental data. A minor amount of research with classical models was in the area of test-score equating. Classical item analysis procedures, however, received little attention. A fair amount of research during the period was devoted to different item types and test item response modes as replacements for the ubiquitous multiple-choice item. Several types of true-false items were proposed, and formula scoring was studied by a number of researchers in an attempt to reduce guessing effects. The perennial topic of response option weighting received attention, with efforts oriented toward demonstrating effects on validity and reliability. Response modes studied included answer-until-correct, confidence weighting, and free-response.

A number of alternatives to classical test theory were studied in an attempt to solve some of the problems for which classical test theory has proven to be inadequate. Research on criterion-referenced testing continued during this period. Latent trait test theory (item response theory, or IRT) received considerable attention. Research on the 1-parameter IRT model continued to address problems of parameter estimation, model fit, and equating. The question of the person-free and sample-free characteristics of this model (i.e., its robustness) were investigated, with results generally supporting these desirable characteristics. In addition, a special case of this model that can account for guessing was developed, and the model was generalized and successfully applied to polychotomous attitude types of items. Considerable research occurred on the 2- and 3-parameter IRT models. The concept of information as a replacement for classical reliability concepts was studied, and its uses in developing parallel tests were described. As with the 1-parameter IRT model, problems of parameter estimation and equating were investigated. These IRT models were successfully applied to problems of item option weighting and adaptive testing. Important developments with these models during the period included the demonstration of their relationship with other psychological measurement models, and methods for determining fit of individuals to IRT models. As another alternative to classical test theory, order models were developed and studied, and several other models were proposed.

Validity issues were also studied during this period. A number of approaches to the analysis of multitrait-multimethod matrices were proposed and compared, including some based on structural equations models. Issues of predictive validity studied included necessary sample sizes, validity generalization, and moderator and suppressor effects. Test fairness issues and their effects on validity received considerable attention. Concern was with (1) bias in selection; (2) fairness to minorities, including differential and single-groups validity and comparisons of regression lines, adverse impact, and bias in test content; and (3) fairness to women.

It is concluded that little of consequence was accomplished in classical test theory during this period. The most important developments were in alternatives to classical test theory, primarily item response theory. Research in this area resulted in data and other developments that will permit a better understanding of the range of applicability of these models and their potential for solving measurement problems not solvable by classical models.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

## CONTENTS

Introduction.....	1
Classical Test Theory and Methods.....	1
Reliability.....	1
Reliability Coefficients.....	1
Generalizability Theory.....	4
Other Reliability Issues.....	5
Other Applications.....	6
Item Analysis.....	7
Item Response Modes.....	7
True-False Tests.....	8
Formula Scoring in Multiple-Choice Tests.....	8
Option Weighting.....	9
Different Response Modes.....	10
Alternatives to Classical Test Theory.....	11
Criterion-Referenced Testing.....	12
Conceptual Issues.....	12
Technical Issues.....	13
Latent Trait Test Theory.....	15
1-Parameter Logistic Model.....	16
Estimation, Model Fit and Equating.....	16
Person-Free/Sample-Free Measurement.....	17
Other Developments.....	18
2- and 3-Parameter IRT Models.....	18
Parameter Estimation and Equating.....	19
Applications: Option Weighting, Adaptive/Tailored Testing.....	20
Relationships with Other Psychometric Models.....	21
Person Fit.....	21
Order Models.....	22
Miscellaneous Models.....	23
Validity.....	23
Content and Construct Validity.....	23
Multitrait-Multimethod Matrices.....	24
Predictive Validity.....	25
Moderator and Suppressor Effects.....	26
Educational Applications.....	26
Test Fairness.....	27
Definitions of Test Fairness.....	27
Bias in Selection.....	27
Fairness to Minorities.....	28
Differential and Single-Groups Validity.....	29
Comparisons of Regression Lines.....	30
Adverse Impact.....	30
Bias in Test Content.....	31
Alternatives to Tests.....	31
Fairness to Women.....	31
Summary and Conclusions.....	32
References.....	34

Accession For.....	<input checked="" type="checkbox"/>
NTIS GRA&I.....	<input checked="" type="checkbox"/>
DTIC TAB.....	<input type="checkbox"/>
Unannounced.....	<input type="checkbox"/>
Justification.....	
.....	
Distribution/.....	
Availability Codes	
Avail and/or	
Special	

A

Acknowledgements

This paper is an extended version of a review prepared for the Annual Review of Psychology, Volume 32, pages 629-658. Preparation of this review was partially supported by a National Academy of Education Spencer Fellowship to Mark L. Davison; and by funds to David J. Weiss from the Army Research Institute, the Air Force Human Resources Laboratory, the Air Force Office of Scientific Research, and the Office of Naval Research, and monitored by the Office of Naval Research.

Technical Editor: Barbara Leslie Camm

## REVIEW OF TEST THEORY AND METHODS

This review is concerned with the period January 1975 through December 1979, including a few papers published in early 1980. The primary focus is the published journal literature, although some books, technical reports, and unpublished literature have been included where relevant. The focus of the review is on practical procedures for converting psychological observations into numerical form, commonly referred to as "test theory." Both the theory and the resulting methodologies are reviewed. Excluded are procedures commonly used for attitude scaling, both unidimensional and multidimensional. However, some scaling methods that either have relationships with or utility for testing in the ability and achievement domains have been included, even though they may technically be considered to be scaling methods. Also not included is the considerable literature on data analytic procedures such as factor analysis, multiple regression, most of the literature on structural equations analysis, and statistical procedures, which are considered by some to be part of psychological measurement. The review also does not include the growing literature on problems of reliability of observations (e.g. interrater reliability) or such measurement approaches as functional measurement, which have had little application to the general problem of measuring individual differences. Thus, the review is concerned with procedures for the measurement of ability, aptitude, and other cognitive variables and for problems of estimating the precision and utility (validity) of measurements of this type.

### CLASSICAL TEST THEORY AND METHODS

#### Reliability

##### Reliability Coefficients

Classical test theory (CTT), which has its roots in work by Spearman in the early 1900s, now is approximating its 75th birthday. Despite Lumsden's (1976) critique of CTT, research related to it seems to continue unabated. Perhaps this attests to the usefulness of this approach to instrument construction, or perhaps it attests to the inertia built into a system of education and training that produces researchers who continue to perpetuate methodologies which, although useful, can be replaced by more coherent methodologies.

Research on reliability estimation in CTT continues to focus on minor modifications of old standby coefficients. Thus, Huck (1978a) has modified Hoyt's analysis of variance reliability estimation procedure (originally developed in 1941) to better estimate the "true" reliability coefficient. The result is a higher reliability estimate by better specifying the error variance, which Hoyt originally defined as interaction of persons and items. Kaiser and Michael (1977) show that the "old faithful" Kuder-Richardson Formula 20 (K-R 20) can be estimated from factor scores derived from "Little Jiffy" factor analysis. In two closely related papers, Raju (1977b, 1979) generalizes coefficient alpha (a special case of both K-R 20 and Hoyt's coefficient) to the reliability coefficient for a "test battery." Although the development seems mathematically appropriate, Raju does not attempt to describe what the reliability of the test

battery means; this lack of coherence is characteristic of much of the research on reliability coefficients in CTT. Is a test battery reliable when the tests are all highly intercorrelated? If so, of what use is such a test battery? What kind of standard error of measurement can be derived from the reliability coefficient for a test battery? How would such a standard error of measurement be used and interpreted in any practical situation? These are some of the kinds of questions that need to be considered with regard to the development of such a generalization of coefficient alpha.

CTT psychometricians seem to be playing, "Will the real lower bound please stand up?" This compulsion has been pursued by ten Berge and Zegers (1978), Jackson and Agunwamba (1977), Nicewander (1975), and Woodhouse and Jackson (1977). These papers "improve upon" work done by Guttman in the 1940s by attempting to better estimate the true reliability from a given data set. Nicewander (1975) shows a relationship between image factor analysis and one of Guttman's lower bounds. The search is extended by Jackson (1979) from internal consistency coefficients to split-half coefficients, even though more appropriate methods exist for estimating internal consistency of a set of items. Methods for better estimating split-half reliability were further studied by Callender and Osburn (1977a) who developed an algorithm to generate the maximum split-half reliability for a set of test items. They subsequently (Callender & Osburn 1977b) show that their sample-based maximized split-half routine gives a better estimate of population reliability than do some of the internal consistency methods, but under the unrealistic conditions of tau equivalence, i.e. a linear relation between true scores on the two halves.

Of course, the Spearman-Brown (S-B) formula, now well into middle age, is still a topic of research in CTT. Allison (1975) generalizes the formula to fractional length tests, and Feldt (1975) provides a formula for the situation in which the assumption of equal variances is not met. Similar to Callender and Osburn's (1977a, 1977b) approach, Feldt's coefficient makes the relatively strong assumption that true scores on the two subtests are perfectly correlated. Another sample-based optimization procedure was presented by Huck (1978b) in his solution to the problem of estimating reliability when items are equally difficult. The only really "new" reliability estimation procedure to appear during this period is a maximum likelihood factor analytic method developed by Werts, Rock, Linn, and Jöreskog (1978).

Common to all these reliability estimation procedures, however, is a major weakness of CTT--the sample-based nature of all of the estimation procedures used for reliability. Thus, reliability estimates are specifically a function of the particular set of items and particular sample of individuals on which the data have been collected. The logical fallacy, of course, is that the translation of reliability coefficients into errors of measurement results in errors of measurement that are specific to a particular test administration event. Consequently, the same individual tested with two different groups of individuals may obtain two different errors of measurement and estimates of true score, based simply on the group of testees with which the person has been tested. This is a serious problem in CTT which cannot adequately be solved by sample-based methods for determining test scores or estimates of precision of measurement.

Nevertheless, CTT marches on. Much of the reliability research continues to concentrate on coefficient alpha, with a recent salutary trend toward methods

for testing the significance of alpha or testing the difference in alpha coefficients from different groups. Pandey and Hubert (1975) compare several interval estimation procedures for coefficient alpha, and Joe and Woodward (1975) provide an approximate confidence interval for maximum coefficient alpha, based on work by Lord (1958), which showed maximum alpha to be a function of the item inter-correlations for a set of test items. Woodward and Bentler (1978) provide a statistical lower bound for population reliability that is useful in estimating population values of reliability from sample estimates, which are usually higher due to sampling error. They use the sampling distribution of estimated alpha coefficients to obtain a new coefficient which better estimates the population reliability. Two new reliability coefficients and one old one are estimated by Sedere and Feldt (1977) in comparison to the theoretical distribution of alpha; these authors define conditions under which each of the estimates of reliability studied appears to be appropriate.

One of the most useful developments during this period is a test for alpha coefficients on independent samples (Hakstian & Whalen 1976) that is useful for comparing the alpha coefficients derived from different groups of individuals, such as individuals in different treatment conditions. Their development is supported by simulation data, and is useful since it allows conclusions to be drawn about the effects of testing conditions on measurement precision, an area of research which has not received much attention in past years. Thus, although many authors have hypothesized the effects of testing conditions on the precision, or reliability, of measurement, the existence of a statistical test for independent samples to compare such reliability coefficients is a useful development.

Another major problem with CTT has been in the confusion that it has engendered among the concepts of internal consistency, homogeneity, and unidimensionality. This is exemplified by the paper by Green, Lissitz, and Mulaik (1977), in which homogeneity and unidimensionality are equated as follows: "Homogeneous items have but a single common factor among them and are related to the underlying factor of ability or attitude in a linear matter" (p. 830), whereas internal consistency is defined as "interrelatedness but not necessarily unidimensionality." A related article by Terwilliger and Lele (1977) attempts to clarify the relationships among internal consistency, homogeneity, and Guttman's idea of reproducibility. When considered together, these articles stand in sharp contrast to each other due to the serious confusion concerning these concepts that has developed in the reliability literature. The confusion is exemplified by Green et al.'s (1977) equating of homogeneity and unidimensionality, whereas Terwilliger and Lele's (1977) use of homogeneity clarifies one use of the term.

To avoid perpetuating confusing terminology, homogeneity should be used only in the sense referred to by Loevinger (1957), rather than in the sense used by Green et al. (1977). Internal consistency is reflected by the alpha coefficient and its derivatives; it refers to the degree of average item intercorrelation among a set of test items. A set of items is internally consistent when the average intercorrelation is high; it is not internally consistent when the average item intercorrelation is low. Linear unidimensionality is what Green et al. have called homogeneity. Linear unidimensionality means a single common linear factor; and if the factor is prominent, a unidimensional set of items will also be internally consistent.

Homogeneity, on the other hand, is not unidimensionality. Homogeneity (in Loevinger's, 1957, sense) relates to the ratio of the sum of the item covariances to the maximum item covariance. In the extreme, when the sum of the item covariances is equal to the maximum possible item covariance, a linearly unidimensional set of items might result. However, homogeneity (in Loevinger's sense) does not index linear unidimensionality except at that positive extreme. Linear unidimensionality is indexed by the lack of variance of the inter-item correlations and by a relatively high mean value of item intercorrelation, which will result in a single common factor. Item intercorrelations can be high, on the average, yet have substantial variance. In that case, more than one factor may exist in the item intercorrelation matrix. Consequently, the use of the term homogeneous should not be equated with the term unidimensional. Rather, homogeneous should be used only in the sense defined by Loevinger, and "internally consistent" (referring to the degree of average item intercorrelation) should be used instead. When linear unidimensionality is explicitly assumed, that term should be used rather than other terms which have other meanings.

Somewhere during the three-quarter century history of CTT, the major purpose of reliability estimation seems to have been lost. Reliability coefficients in and of themselves have little utility for practical situations, except for comparing their magnitudes in order to evaluate measuring instruments. However, every reliability coefficient should be viewed only as a step toward estimating the precision of an individual score. In the history of scientific investigation, only psychometrics has developed the concept of reliability coefficients. In all other applications of measurement, e.g. physics and other sciences, precision of measurement is indexed by the probable deviation of an observed value from some true value. Alternatively, precision is estimated by some confidence interval that is likely to include the true value. Thus, measurement of height is accurate to plus or minus some degree of error. Yet the preoccupation in psychometrics seems to be that of estimating reliability coefficients, with little attention paid to the problem of estimating the precision of an individual measurement or, conversely, the error of measurement.

In the period under review, only two papers have been concerned with the standard error of measurement, the psychometric analogue to physical errors of measurement. Dudek (1979) revived some long-forgotten history of interpretations concerning the standard error of measurement, depending on whether the user of the measurement is concerned with estimating true score, or placing an error band around an observed score. Kleinke (1979) demonstrates bias in some approximations to the standard error of measurement based on reliability coefficients, and Whitely (1979) is concerned with methods for estimating measurement error on highly speeded tests, an issue that has not been adequately resolved previously within the context of CTT.

### Generalizability Theory

Although not technically a part of CTT, generalizability theory is really only a generalization of Hoyt's basic idea of variance decomposition of a person by items response matrix, originally proposed in 1941. It is also heavily rooted in "domain sampling" theory, which was developed most explicitly by Tryon (1957). Although originally proposed by Cronbach, Gleser, Nanda, and Rajaratnam in 1972, because of its complexity and the lack of procedures for estimating many of its parameters, generalizability theory had not been brought to practi-

cal status prior to the period under review. During this period, several developments in generalizability theory have occurred.

Kaiser and Michael (1975) derived Tryon's (1957) domain validity coefficient (which bears some striking similarities to Cronbach et al.'s (1972) generalizability coefficient) using minimal assumptions. It, of course, turned out to be a generalized version of the alpha coefficient, thus perpetuating what they characterize as "one of the favorite indoor sports of psychometricians" (p. 34), but requires no assumptions about the means, variances, covariances, or structure of the items. McDonald (1978) draws relationships between the idea of "domain validity" and the concepts of generalizability theory, whereas Cardinet, Tourneur, and Allal (1976) criticize applications of generalizability theory to educational measurement and suggest examples of situations in which the variables on which differentiation is desired are opposite those that are appropriate for typical generalizability analyses. Joe and Woodward (1976) develop multivariate generalizability theory, estimating components of maximum generalizability and multifacet experimental designs with multiple dependent variables (which turn out to be multivariate extensions of the Spearman-Brown formula). Brennan attempts to bring generalizability theory to the user (e.g. Brennan 1980a), develops algorithms and procedures for the estimation of variance components (e.g. Brennan 1975) and provides computer programs for implementing aspects of generalizability theory (Brennan 1980b).

Although generalizability theory appears to be a useful conceptualization that has begun to reach practitioners for practical application, potential users should carefully consider some of its assumptions before becoming too enamored with it. Rozeboom (1978) criticized both Kaiser and Michael (1975) and generalizability theory in terms of the conceptual existence of a domain. Rozeboom describes the logical impossibility of sampling from a domain in order to make the assumptions necessary to generate both coefficient alpha and generalizability theory. He also indicates that domain validity provides no information about the domain, since it is strictly a function of the number of items, noting further that domains are likely to be multidimensional, and that only the first dimension is estimated by domain validity and the variance components of generalizability theory. Thus, Rozeboom questions the implicit and explicit assumptions of generalizability theory and its predecessors, with some cogent criticisms which should be carefully considered by persons who use this approach to the estimation of measurement precision.

#### Other Reliability Issues

The measurement of change has received a fair amount of attention during the period reviewed. A minor controversy arose between Overall and Woodward (1975, 1976) and Fleiss (1976) when Overall and Woodward demonstrated by some derivations that the power of  $t$ -tests is at the maximum when the reliability of difference scores is 0. Fleiss showed that Overall and Woodward assumed a restrictive model with no interaction between subjects and time, but when a less restrictive (and more realistic) model was assumed, then the maximum power of the  $t$ -test for correlated measures is attained when the reliability of difference scores is a maximum. Overall and Woodward (1976) replied that Fleiss was concerned with the reliability in the original pre-test and post-test scores, and that his findings are correct for that situation, but that he did not consider the reliability of difference scores. Overall and Woodward then reas-

serted that the power of a pre-post-test  $t$ -test is highest when difference scores are unreliable. Williams and Zimmerman (1977) discuss the reliability of difference scores when errors are correlated and conclude that when errors are correlated (which may well be the case in a number of applications), difference scores can be more reliable when they are not correlated, which is the usual assumption made.

Other problems in the measurement of change are addressed by Bond (1979), Cascio and Kurtines (1977), Corder-Bolz (1978), Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979), Hoogstraten (1979), Linn and Slinde (1977), and Richards (1975). Werts, Linn, and Jöreskog (1977) present a maximum likelihood factor model for estimating reliabilities in unattenuated correlations between growth measures, whereas Werts and Hilton (1977) and Hilton (1976)--also using structural equations models--describe direct estimates of change score reliabilities and unattenuated correlations between pre-test and change scores.

Applications of reliability theory in experimental design and the analysis of experimental data have begun to receive some attention. Nicewander and Price (1978) extend the Overall and Woodward (1975, 1976) and Fleiss (1976) controversy to a discussion of the reliability of dependent variables and the power of significance tests. They indicate that reliability is not related to power for controlled experiments and that under certain conditions both of the previous authors are correct. Their discussion centers on the problem of an individual differences versus an experimental focus in the research design, since considerations of both between subjects sampling variance and measurement error variance are relevant to the reliability and power issue. Subkoviak and Levin (1977; Levin & Subkoviak 1977, 1978) and Forsyth (1978a, 1978b) discuss the effects of measurement errors on the power of statistical tests. Careful reading of this interchange indicates that the nature of the experimental design plays some role in the effect of reliability on power, as does whether observed scores or true scores are being considered.

#### Other Applications

A few other minor issues in GTT were also studied during this period. Slinde and Linn (1977b) compared linear versus equipercentile methods for equating different tests given to different groups of individuals. They found that the equipercentile method was better than the linear method but that both had some serious problems in properly equating test scores. Rubin and Thayer (1978) also considered the problem of test equating in the situation where a reference test is given to each of a number of groups and new tests are given to only one of the groups. The problem they considered was to estimate scores on the new tests even though everyone did not take them. Their method is limited, however, to the use of "plausible" values estimated for the intercorrelations among the new tests and the standard reference test. Healy (1979) formulated a test of the linear relation between two true scores, but the test is limited to the situation in which the covariance matrices of the two tests are equal. Lord and Stocking (1976) developed a method for estimating the regression function of true score and observed score assuming a binomial error model but not assuming that true score and error scores are linearly related. The attenuation paradox, and its relationship to the distribution of ability, is considered by Nicewander, Price, Mendoza, and Henderson (1977); they indicate that the attenuation paradox will result, regardless of the distribution of ability, if items of

"perfect discrimination" are used. Finally, Zimmerman (1976) develops CTT from concepts of probability theory and statistical sampling theory, rather than from the usual assumptions. The problem, of course, is that neither probability theory nor statistical sampling theory have any relationship to the psychological processes underlying test behavior, thus further removing CTT from the mainstream of psychology.

### Item Analysis

Virtually no progress was made during the period in the area of item analysis. The papers that have appeared seem to be either repetitions of what has been done for years in item analysis or minor extensions of techniques already available. For example, D'Agostino and Cureton (1975) concluded that the old "27% rule"--contrasting the proportions correct for the upper and lower 27% of the score distribution--is acceptable but that a 21% rule would be better. Berk (1978) empirically evaluated formulas for corrections of item-total point-biserial correlations, and Beuchert and Mendoza (1979) find very few differences among 10 item discrimination indices, as did Oosterhof (1977) in his factor analysis of 19 item discrimination indices. Both Aiken (1979) and Hoffman (1975) concerned themselves with the age old issue of choosing items based on both difficulty and discrimination indices.

In the area of personality measurement, Neill and Jackson (1976) developed an item efficiency index designed to reduce scale intercorrelations for a multiscale measure, since multiscale measures are likely to be more valid against external criteria if interscale correlations are low. Their method is illustrated with personality data but has applications to other multiscale batteries. Yet another demonstration that observed score distributions, and therefore proportion of testees passing any given cutoff score, can be manipulated by the way items are selected on item difficulty is given by Dyck and Poencke-Schuyten (1976); Nevo (1977) demonstrates what should be obvious--that traditional item analysis does not increase test-retest reliability; and in a somewhat useful development, Hsu (1978) gives appropriate alpha levels to use in testing item analysis statistics when the use of multiple items changes the experimentwise error rate.

### Item Response Modes

A little psychology begins to interact with test theory when real people begin to take real test items. Since the beginning of CTT, the "objective" test item has been the rule, usually characterized as a dichotomous (true-false) item or as the ubiquitous multiple-choice item. Ever since the invention of these test item formats, a number of questions have arisen, and research still continues in an attempt to answer them. The questions arise from the fact that the objective test format leads to some loss in information on a testee's ability/achievement level and may introduce other sources of variability in test item responses (such as guessing) in addition to the variable that the test item is designed to measure. Research on these issues has consistently manifested itself in several areas: (1) attempts to study the effects of guessing on various item response formats; (2) attempts to reduce the effects of guessing by the use of various scoring formulas; (3) the effect of various item option weighting schemes; (4) the use of alternative response formats within multiple-choice items; and (5) studies on the use of alternative response formats.

True-false tests. Psychometricians frequently decry the use of true-false (T-F) tests probably because of the high probability of guessing on these kinds of items. Consequently, a number of authors have proposed alternatives to T-F tests such as paired T-F tests (Eakin & Long 1977) in which two T-F items are presented together and the testee answers as if the item were a four-choice multiple-choice item, indicating the truth or falseness of each of the four item pairs. These authors suggest that this approach better reflects the true knowledge, in comparison to a number-correct score (but not in comparison to a correct minus incorrect score). Hsu (1979b) studied a similar type of T-F test and found an interaction with knowledge level of examinees, in which the grouped T-F items (with two or three items per cluster) were better for low ability testees than the separate T-F items, which were better for moderate and high ability testees. Hsu (1979b) followed up on Eakin and Long's results and reported that their method of scoring the paired-item T-F test results in misranking testees under certain conditions. Finally, Aiken and Williams (1978) studied the effects of instructions to testees on seven methods of scoring T-F items. Their results supported those of Hsu (1979a) showing an interaction of scoring methods with ability levels.

Formula scoring in multiple-choice tests. Given the decision to use a multiple-choice (M-C) item, the psychometrician must next determine how that test is to be scored. Obviously, it is simple to count the number of items answered correctly, but the problem of guessing, when a testee does not know the answer, rears its ugly head. The history of CTT has concerned itself for some time with the guessing problem, and one classical solution is that of formula scoring. A formula score is a score other than a number-correct score that is designed to adjust the number-correct score for chance successes due to guessing. A typical formula score is the number correct minus some fraction of the number answered incorrectly. Lord (1975a, 1975b) began a minor controversy by studying the relative efficiency of number-correct and formula scores, using concepts of latent trait test theory to compare the two scoring methods. Although this approach may be unfamiliar to many, since scoring methods are usually compared in terms of reliability, the comparison made by Lord can be interpreted in a quasireliability sense through the relationships between the concept of information and error of measurement (see below). Lord's main result was that formula scoring is 3% to 9% more efficient than number-correct scoring, primarily for testees of moderate to low ability levels.

A number of researchers have taken issue with Lord's finding, based primarily on the particular set of assumptions that Lord implicitly made in his study. Specifically, Lord assumed that formula scoring instructions concerning omitting responses would result in random guesses if the same test was administered under number-correct instructions. Cross and Frary (1977) administered a 20-item test, including four items with no correct answer, using formula scoring instructions (as specified by Lord 1975a) then had testees answer the omitted items in a different color. For Cross and Frary's (1977) more than 407 testees, results showed that there were many omitted items that the testees had a better than chance probability of answering correctly. The results also indicated that 27% of the testees did not recall the exact directions under which the test was administered. Their results did not support Lord's assumptions however, even for those who did understand and comply with the directions. Cross and Frary concluded, therefore, that the correction for guessing should not be used. Similarly, Wood (1976) administered M-C tests under three different kinds of in-

structions and examined whether the instructions produced the desired effects for different groups. Wood's conclusion is similar to Cross and Frary's in that he found the instructions worked only to some extent in eliminating blind guessing.

Similarly, Rowley and Traub (1977) also studied the behavior of testees under Lord's (1975a) formula scoring instructions in an ability testing situation. Rowley and Traub's (1977) conclusion is that formula scoring is not logically supported because of differential personality factors on the part of the testees, which enter into the responses to M-C test items. This result is supported by Slatker, Crehan, and Koehler (1975), who examined risk-taking on objective examinations and found that guessing occurred even when the penalty for guessing was known. They also found age and sex differences in means and stabilities of a risk-taking score. In an attempt to develop better formula scores, Reid (1977) and Abu-Sayf (1977) developed new versions of formula scores, and Molenaar (1977) took a Bayesian approach to correcting for random guessing. The weight of the data accumulated during this period is that formula scoring of M-C tests has some problems but that number correct scores are also nonoptimal when guessing is likely to occur. Finally, Grier (1975) suggested that three-alternative M-C items (scored by number correct) maximize expected test reliability, with two-alternative items better than four-, five-, or six-alternative items. It should be noted, however, that this conclusion is valid only if test length is increased to compensate so that the number of alternatives times the number of items is fixed. His results are also restricted to the specified assumptions regarding the quality of the test items and should not be accepted as an across-the-board recommendation.

Option weighting. Since formula scoring does little to increase the amount of information obtainable due to partial knowledge from M-C items, a considerable amount of research over the years has been spent in comparing methods for weighting the options of M-C items and including those option weights in total scores. The last five-year period was no exception to this trend. Building on work by Mosier in the 1930s and Guttman in the 1940s, methods of option weighting have continued to receive considerable amounts of attention. Claudy (1978) proposed the biserial correlation of each option with total score as an item option weighting scheme. He compared his method with a number of other option-weighting schemes and evaluated the results with split-half reliability coefficients. Serlin and Kaiser (1978) increased the reliability of M-C tests by weighting item responses based on their loadings on the first principal component of the intercorrelations among the 0-1 weights for each item alternative. They show increases in alpha from .44 to .77 for 0-1 scoring but did not cross-validate their results. Downey (1979), Raffeld (1975), Reiley (1975), and Echternacht (1976) compared various item option weighting schemes in terms of reliability and validity against a number of criteria. Their general conclusion is that there are almost always increases in reliability for most item option weighting schemes and occasional increases in validity, depending on the nature of the validity criterion used. Bejar and Weiss (1977) demonstrated that the increases in the reliability of option-weighting schemes are generally a function of the item intercorrelations, whereas Echternacht (1975) derived formulas for estimating the variances of one type of option weights and provided suggestions for item writing to decrease these variances. Cross and Frary (1978) studied the effects of "guess" and "do not guess" instructions on empirical choice weighting and compared them to number-correct and formula scores. Their

results showed no differences in validities between the guessing conditions but higher validities for the empirical choice weighting procedure.

Different response modes. Since formula scoring offers little improvement over number-correct scoring, and the results regarding differential option weighting have been mixed, the search continues for ways to improve the characteristics of scores obtained from M-C test items. Dunkin and Milton (1978) propose simply modifying the answering procedure in a M-C item by constructing items of any number of correct responses and by permitting the testee to choose any number of answers as correct. They call these multiple-answer-multiple-choice items, develop relevant scoring rules, and evaluate some Bayesian and minimax strategies for responding to these kinds of items. Their proposal is interesting and might stimulate some relevant research that might result in improved scores. They also examine the possibility of probabilistic responding in which the testee responds with subjective probabilities for each alternative, one at a time.

Another procedure for a different kind of responding to M-C items is the answer-until-correct (AUC) procedure. These procedures have been studied since the early 1950s, based on work by Coombs (1953). Recent research on this problem illustrates theoretical increases in efficiency for this procedure (Gibbons, Olkin, & Sobel 1979), provides equations making it possible to select items that minimize guessing under AUC administration (Kane & Moloney 1978), and demonstrates higher levels of reliability but lower levels of validity (Hanna 1975). Although AUC procedures may reduce the number of items required in a test, their use may also increase testing time.

Confidence weighting also continues to be studied. In this procedure, testees answer an item either (1) by choosing a correct answer and assigning some confidence level to their choice or (2) by distributing some fixed number of points (e.g. 100), indicating their confidence that each of the M-C alternatives is correct. Another variation is simply ordering the alternatives in the item on the basis of correctness. Diamond (1975), Poizner, Nicewander, and Gettys (1978), Wen (1975), and Pugh and Brunza (1975) all found increases in reliability of confidence-weighted scores over nonconfidence-weighted scores, and Abu-sayf and Diamond (1976) found increases both in reliability and validity.

In a relatively comprehensive comparison of several methods of assessing partial knowledge in M-C items, Hakstian and Kansup (1975) and Kansup and Hakstian (1975) compared both conventional and confidence rating instructional sets and a number of different scoring methods--including confidence rating scoring on both verbal ability and mathematical reading tests--in terms of both reliability and validity. They concluded that confidence-rated scores were more internally consistent and more stable than conventional scores. They also found higher validities for confidence-rated mathematics scores in comparison to the conventional scores, but the differences were not statistically significant. For the verbal ability tests, the conventional scores were more valid than any of the option-rating scores. They conclude that although the validity results were mixed, including some higher levels of validity for confidence-rated scores on their grade-point average criterion, confidence rating testing time was longer, and more conventional test items could compensate. However, these are simply extrapolations from their data that have not been supported by empirical results. Thus, the data tended to show some specific utility for the confi-

dence-rating approaches, perhaps greatest in more complex abilities such as mathematics than in abilities that rely less on information-processing characteristics of the individual, such as verbal ability.

Linn (1976b) critically evaluated probability responding procedures, especially the "value assigned" score, where a person's score is the probability value assigned to correct responses. Linn indicated that the value-assigned score does not provide a reproducing scoring system, since the best strategy is an all or none response; and since testees do not respond rationally to differing probability situations (where rationally is defined as maximizing the total scores), probabilistic testing procedures must be designed in terms of the instructions to the testee in order to eliminate these problems. Linn's suggestion is that if probabilistic procedures are combined with appropriate instructions to the testee, gains in utility of the test scores might result.

One obvious solution to the problems inherent in M-C tests is simply to replace the M-C item with some other kind of test item. Obviously, the M-C item has retained its popularity because of the ease in objective scoring of such items and the rapidity with which scores can be derived. As recent research on computer-administered testing (e.g. Brown & Weiss 1977; Cory 1976; Cory, Rimland & Bryson 1977; Kingsbury & Weiss 1979a; Weiss 1976) results in the ultimate replacement of the paper and pencil test by computer-administered tests, the possibility of replacing the M-C item with free-response items becomes realistic. Very little research, however, has been done on the use of free-response items. Vale and Weiss (1977) compared free-response and M-C vocabulary tests in terms of information concepts from latent trait test theory. Their data show substantial increases in precision of measurement to be gained from the use of free-response items in comparison to M-C items, primarily for testees of middle and high ability levels.

Traub and Fisher (1977) compared the factor structure of free-response and M-C items, as well as AUC scoring, using verbal comprehension and mathematical reasoning items. Their tests were carefully designed to be equivalent, and the factor structures were compared by methods of confirmatory factor analysis. The results showed that the different item formats measured the same factors in the mathematics test but that the structure of the verbal test was a function of the format, with the free-response format resulting in a more complex factorial structure than the other two item types. These results suggest that additional information may exist in free-response data, which may permit different kinds of measurement to be obtained from different response formats, and that care should be taken in the translation of existing tests into new response formats, since different dimensions may be measured by the same items cast in different response formats. Harris and Pearlman (1978) provide a domain-oriented index of response agreement for free-response items.

#### ALTERNATIVES TO CLASSICAL TEST THEORY

Because classical test theory has been unable to adequately solve a number of testing problems during its history, several alternative test models have been proposed. Criterion-referenced testing has developed and flourishes in an attempt to solve the mastery testing problem. Latent-trait based test theories continue to be refined and applied to a wide range of problems for which classi-

cal test theory is inadequate. Order-based test models, which have developed during the last few years, show some promise for measuring certain kinds of psychological variables, and a few other miscellaneous new approaches have been proposed.

### Criterion-Referenced Testing

The extensive literature on criterion-referenced testing can be divided into two parts: (1) articles dealing with conceptual issues and (2) articles dealing with technical considerations.

#### Conceptual Issues

Popham (1975, p. 130) and Hambleton, Swaminathan, Algina, and Coulson (1978, p. 2) define a criterion-referenced test (CRT) as one "used to ascertain an individual's status [the individual's domain score] with respect to a well-defined behavior domain." If this definition captured the sole essence of CRT, the term CRT would be less appropriate than Hively's (Hively, Patterson, & Page 1968) "domain-referenced testing," Ebel's (1962) "content-referenced testing," or Osburn's (1968) "universe-defined testing." Swaminathan, Hambleton, and Algina (1974) pointed out that a CRT is used primarily to ascertain a student's standing with respect to a prescribed mastery (pass-fail) standard and hence the name CRT. In some cases, prior normative information is used in setting the criterion (Pinney 1979; Popham 1978), a practice that blurs the distinction between norm- and criterion-referenced testing but that helps avoid unrealistically high or low criteria.

Glass (1978), along with Burton (1978), Levin (1978), Linn (1978a), and Messick (1975), criticized the use of mastery criteria in testing because such criteria are necessarily arbitrary. Linn (1978a), however, doubts that educators can easily sidestep the demand for standards reflected in opinion polls and deliberations of state legislatures. Glass (1978) concludes that measurements should be based on comparative standards of better or worse rather than on arbitrary mastery levels.

In a well-reasoned response to Glass (1978), Hambleton (1978) argued that mastery criteria are arbitrary, but in the best sense of the word; they are standards reflecting professional judgment and discretion. For all their faults, he argues, such criteria are still the best basis for educational decisions. Block (1978), Popham (1978), and Scriven (1978) offered further rebuttals to the critics of standard setting, and Terwilliger (1977) discussed the philosophical issues involved in setting standards.

When institutional limitations impose constraints on the number of examinees who can be selected, these constraints often imply a normative approach. For instance, if a college has facilities for only 500 new freshmen and wants the 500 that are best qualified, then selection will necessarily be based on normative comparisons between applicants. Often, however, no such constraints exist. For instance, nothing typically constrains the number of students receiving a passing course grade, the number of students moved to the next level of instruction in computer-aided instruction, the number of drivers' licenses issued by a state, or the number of professionals licensed by a state for a given field. In situations where no institutional constraints operate, the use of

mastery criteria is commonplace, sensible, and not overly controversial.

The controversy over standards centers primarily around their use in large-scale statewide or districtwide assessments to determine who will and will not be given diplomas. Because any adopted standard can be used either as a guideline in evaluating student performance or as a tool for assigning diplomas, the standards debate should center not only around the question of whether standards will be used but also around the question of how they will be used.

#### Technical Issues

A number of technical issues have received minor attention. Wilcox (1976, 1979a) describes methods of deciding the optimal length for a CRT. Both Kingsbury and Weiss (1979) and Spinetti and Hambleton (1977) present computerized adaptive CRT strategies, which can reduce the number of items needed to make mastery decisions. Van der Linden (1979) argues that binomial test models, the models on which much of CRT theory rests, impose some unrealistic conditions on item characteristic curves. Lewis, Wang, and Novick (1975), and Wilcox (1978a, 1979b, 1979c) propose methods of estimating true domain scores on CRTs.

Most of the CRT item analysis procedures are variations of conventional methods (Haladyna 1974; Lord 1977d; Mehrens & Lehman 1978; Panell & Laabs 1979). For instance, Mehrens and Lehman (1978, p. 334) suggest pruning items on the basis of a discrimination index reflecting the difference between the item's difficulty in a pre- and post-instructional group. Several authors argue, however, that pruning items on the basis of difficulty or discrimination indices violates the very concept of a CRT, defined as a test designed to assess an individual's status in a well-defined behavior domain (Levine 1976; Osburn 1968; Shoemaker 1974). Kwansa (1974) found that after items were pruned on the basis of conventional item statistics, the items remaining were not representative of the original domain. With the exception of Rovinelli and Hambleton (1977), the CRT critics of conventional item development procedures have been too glib about the problems in selecting items to be representative of a domain, problems exacerbated by the fact that the domain is often so vaguely defined.

One of the most frequently discussed problems in CRT is that of setting the criterion. Meskauskas (1976) reviews the suggested methods, covering several papers that appeared before the period of this review. Methods of setting the criterion that maximize subjective expected utility functions under various sets of assumptions have been proposed by Macready and Dayton (1977), Huynh (1976b), Huynh and Perney (1979) and Wilcox (1979d). Swaminathan, Hambleton, and Algina (1975), and Wilcox (1977, 1979f) discuss a related problem, that of estimating the probability of false negative and false positive errors in making mastery decisions. Some decision makers, however, may feel uncomfortable making the utility judgments that these criterion-setting methods require. Further, these methods for setting an observed mastery score require that there already exist a set criterion score either on a referral task or in terms of true scores, a requirement that begs the fundamental question in most cases.

Because criterion-referenced tests need not have variance, Millman and Popham (1974) argue that classical reliability statistics are not appropriate measures of test precision. Like Woodson (1974), we are left puzzled. If a major purpose of a CRT is to classify examinees into mastery states (Swaminathan et

al. 1974), then what purpose could be served by a test on which every examinee in the population of examinees for which the test was intended would obtain the same score?

Whether prompted by Millman and Popham (1974) or not, considerable effort has been expended to find alternatives to classical reliability for CRT. Several suggestions are based on a redefinition of test variance as variance around the criterion score, rather than variance around the mean score. Livingston (1972) redefined reliability as the ratio of true score variance about the criterion to observed score variance about the criterion. Brennan and Kane (1977) and Lovett (1977) develop analysis of variance estimates of generalizability and reliability coefficients that are extensions of Livingston's (1972) notions.

Beginning with Swaminathan et al. (1974), a number of coefficients of precision for CRT based on decision concepts have been proposed. Swaminathan et al. proposed Cohen's (1960) kappa, which measures the consistency of decisions on two parallel tests. This index can be estimated directly only if there are two test administrations. Huynh (1976a, 1979), Marshall and Haertel (1976), Strasler and Raeth (1977), and Subkoviak (1976) discussed methods of estimating kappa from a single test administration. Algina and Noe (1978) and Subkoviak (1978) critically evaluated some of these single administration estimates.

Beginning with van der Linden and Mellenbergh (1978), the literature on decision-based coefficients of test precision starts winding toward a surprising conclusion. Mellenbergh and van der Linden (1979) argue that tests should be evaluated, not on the consistency of decisions across two occasions, but on the consistency between decisions based on the test and decisions which would be made if the true scores were known. With this consideration in mind, van der Linden and Mellenbergh (1978), Mellenbergh and van der Linden (1979), and Wilcox (1978b) propose a rescaling of Bayes risk as a decision-theoretic index of test quality. Bayes risk is the expected value of (decision) losses with respect to the joint distribution of random variables  $T$  (true score) and  $X$  (observed score) in a given population. Another decision-theoretic proposal is offered by Livingston and Wingersky (1979). Seemingly to their surprise and ours, van der Linden and Mellenbergh (1978) manage to show that their rescaling of Bayes risk is equal to the classical reliability coefficient if a linear or squared error loss function is assumed and if a linear regression of true on observed scores is assumed.

As van der Linden and Mellenbergh (1978) suggest, a measure of test precision should reflect the correspondence between decisions reached using true and observed scores. Coefficient kappa does not do so. Kappa and coefficients derived from Livingston's work are highly situation specific, because they can vary a great deal depending on where the user sets the criterion. If the premise is accepted that a useful CRT must have nonzero variance in the population for which it is intended, then the classical reliability coefficient may be a suitable index for CRT after all. It has a decision-theoretic interpretation, it does not depend on where the criterion is set, and it is readily understood by many users. The standard error of measurement is also a useful index of CRT precision because it can be used to estimate the probability of misclassifying an examinee for any desired criterion level.

Forsyth (1976), Pandey and Shoemaker (1975), and Raju (1977a) discuss mul-

multiple-matrix sampling techniques that may be used in CRT. Myerberg (1979) found that stratifying items by difficulty and content did not improve estimates of mean test scores in a multiple matrix sampling design.

CRTs have been well represented in classrooms, if not in measurement texts, for as long as there has been education. CRT will continue to endure in the form of classroom exams, licensing exams, and tools of computer-aided instruction. Many of the recently developed psychometric methods for CRT may not endure so long. For example, because they require a previously set criterion score on a referral task or on the true score continuum, the methods of setting the criterion that are based on subjective expected utility theory largely beg the question. Although kappa is gaining in popularity as a measure of test stability, it will not soon supplant standard variance-based indices of test precision. Sophisticated CRT methods of estimating true scores and setting test length can be expected to receive no more use than they have received in more conventional testing.

#### Latent Trait Test Theory

Latent trait test theories have their roots in work in the 1940s by Mosier, Guttman, and Lazarsfeld, among others. Although the basic ideas were known about 40 years ago, the methods could not be successfully applied until high-speed computing equipment was available to psychometricians for research and applications. As this equipment became better available to solve some of the problems of latent trait test theories during the early 1960s, models were further developed and techniques further refined. A second barrier to the application of these techniques was that of their sophisticated mathematical requirements, which barred a number of psychometrically oriented researchers from thoroughly understanding the methods.

Latent trait test theories have been applied and developed under several rubrics. Most well known are item characteristic curve theory, and, more recently, item response theory (IRT). The latter is used here because it emphasizes the psychologically based nature of these theories.

In an attempt to make IRT more useful and more widely understood by practitioners, several articles during the review period have provided a basic introduction to IRT. Hambleton and Cook (1977), in their introductory article to a very useful special issue of the Journal of Educational Measurement, provided a brief introduction to IRT. A more comprehensive and relatively nontechnical review for the uninitiated was provided by Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978). Marco (1977) gives practical examples of applications of IRT, as suggested by Lord (1977c); these include practical examples of designing a multi-purpose test using the information curves of IRT, evaluating a multi-level test, and equating tests on the basis of pre-test statistics.

IRT models are usually differentiated by the number of parameters estimated for the items and the nature of the item characteristic curve or item response function (IRF). IRFs are usually assumed to be either normal or logistic ogives. Since there is a high degree of similarity between the two (although some differences in practical applications; see Kingsbury & Weiss 1979b), that distinction will be ignored here, and the logistic ogive will be assumed. Thus, the basic differentiations among the models are in the number of parameters nec-

essary to describe the shape and location of the IRF. A special case of general IRT, the 1-parameter logistic model, is also known as the Rasch model, having been independently developed by the Danish mathematician. In this model the test items are described only in terms of their difficulties, and their discriminations are assumed to be equal. The usual 2-parameter model (as will be seen, there is a special 2-parameter case of the Rasch model) describes items both by their difficulties and by their discriminations; and the 3-parameter model adds a chance level or a pseudo-guessing parameter as the third descriptor of the IRF.

#### 1-Parameter Logistic Model

Estimation, model fit, and equating. The 1-parameter logistic (1PL) model has generated a substantial amount of research during the review period. As is characteristic of research on IRT models, much of the basic research has been focused on problems of item parameter estimation. Since the 1PL model parameter estimation procedure involves estimating only the difficulty parameters for items along with the ability parameters for individuals, these two parameters are usually estimated simultaneously. However, because of some mathematical problems, they can only be approximated under certain circumstances: Cohen (1979) provides noniterative procedures for estimating ability and difficulty that gives values similar to the maximum likelihood procedures usually used for this process. Wright and Douglas (1977a, 1977b) compare different procedures for estimating these parameters, as do Anderson and Madsen (1977); and Anderson (1977) verifies that the number-correct score is a minimal sufficient statistic in M-C tests for estimating trait levels. He also demonstrates that the number-correct score in the 1PL model is not a function of the item difficulties used in the test, whereas Kearns and Meredith (1975) provide Bayesian procedures for point estimates of 1PL model scores. Their procedure is an empirical Bayes procedure, which like all such procedures is sample dependent and only efficient with large sample sizes.

The problem of parameter estimation in these models relates to the question of fit of the models to data. One important feature of the IRT models, and particularly the 1PL model, is that procedures are available for testing the fit of data to the models. If items do not fit the model, they can be eliminated and a set of items can be identified that do fit the model. When a set of model-fitting items is identified, they permit the use of number-correct score as an indicator of trait level in the 1PL model. However, there has been some question about the utility of tests that do fit the 1PL model.

Wood (1978) fit the 1PL model to simulated coin tosses of 500 subjects on 50 variables, a reasonable ratio of subjects to variables to adequately estimate the parameters of the 1PL model. He found that he was unable to reject 47 of the 50 items for lack of fit at the 95% confidence level of the chi-square test usually used to test the fit of items to the model. Thus, his data suggest that random data would not show substantial nonfit to the model. His data, however, indicated that the discriminations of the items were low and that the ability estimates were essentially the same for all his simulees. His conclusion, however, is that a demonstration of lack of nonfit by itself is not good enough, since most of his randomly derived items fit the model.

The problem here, of course, is that on the basis of a lack of nonfit, some

users of the IPL model would be tempted to conclude that the model does fit the data; item discriminations would be set at 1 and use of the model would continue. However, Wood's data suggests that this would be inappropriate, since the true item discriminations were very low and setting them equal to 1 would result in inappropriate discrimination values and inappropriate indications of the error of measurement for the testees. Thus, additional research is indicated on methods for testing fit of data to the IPL model.

One of the major advantages of IRT models, including the IPL model, is the promise of being able to measure individuals on the same ability scale, regardless of the difficulty of the subset of items on which they are measured. This invariance of ability estimates over item subsets implies the capability of IRT models to equate measurements from different tests, a problem that is not adequately solvable with classical test theory. Thus, the usefulness of the IPL model for vertical equating has been investigated by Slinde and Linn (1977b). Their data suggest problems in the use of the IPL model for vertical equating, since they found mean differences in ability estimates based on high or low ability calibrations in their cross-validation groups, with greater differences for ability levels in the calibration groups that were farther apart. They suggest that perhaps more item parameters are necessary to do a good job of vertical equating.

Gustafsson (1979) suggested that the Slinde and Linn (1979a) results caused a spurious lack of fit to the IPL model by selecting levels of examinee performance on subsets of items only, resulting in a regression effect that could have caused the obtained results. He does admit, however, that there may be problems in vertical equating in the IPL model if guessing exists, since this would introduce a correlation between item difficulty and item discrimination. Slinde and Linn (1979a) analyzed their data in an attempt to eliminate the regression effect suggested by Gustafsson (1979), using a different data set. Their results supported their earlier data indicating problems in vertical equating with the IPL model since the item parameters estimated on high and low groups resulted in different mean ability estimates. The problems were mainly characteristic of the low ability group and therefore may be due to guessing, which is supported by a negative correlation between the difficulties and the discriminations in their low ability group. They do, however, concede that the IPL model may be useful for equating in less extreme situations than used in their data. This is confirmed by their later results (Slinde & Linn 1979b), which support the use of the IPL model in vertical equating for relatively contiguous ability levels but not for those which were further apart. Rentz and Bashaw (1977) also illustrate the use of the IPL model for equating using a linking test of common items.

Person-free/sample-free measurement. Because the IPL IRT model promises measurement that is free of the influence of a specific group of testees or a specific subset of test items, in contrast to the sample-specific measurement of CTT, considerable research continues on these capabilities of the model, independent of the equating problem. Tinsley and Dawis (1975) found the IPL easiness parameters and ability estimates to be invariant over samples of testees differing in ability level for tests of 25 or more items. However, their results indicated that the easiness parameter estimates were no more invariant than the  $z$ -transformed proportion-correct difficulty values. Their data (Tinsley & Dawis 1977) also support the test-free characteristic of the IPL model in that ability estimates for individuals did not differ substantially when they

were based on item subsets of different difficulty levels selected from the same tests. Their data also indicate that the IPL model fit M-C items (but recall Wood's, 1978, study of the chi-square test of fit) after testees with scores in the guessing range (10% to 15% of the sample) were eliminated.

Dinero and Haertl (1977) studied the applicability of the IPL model when item discriminations varied, generating testee response data from the 3-parameter model and then fitting the IP model. The results indicated that when the distribution of item discriminations was uniform, the IP model did not fit the data, but when there were substantial numbers of items with similar discriminations (normal or skewed distributions of discriminations), the fit of the model was good. Again, these fit studies should be interpreted with regard to Wood's (1978) study of the test of fit. Whitely and Dawis (1976) found IPL item difficulty parameter estimates to differ as a function of test context, thereby questioning the invariance characteristics of the IPL model. Whitely (1977), in response to the paper by Wright (1977a), which gives some insights into some aspects of the IPL model, agrees with Wood's (1978) later conclusion that the chi-square test of the fit of the model has little power for small samples and does not do well for sample sizes even up to 800.

Thus, the data on the robustness of the IPL model are equivocal, since some studies support invariance of ability estimates and item parameter estimates over item and person sampling, whereas others suggest that the invariance may not be as great as promised by the model. The interpretation of these results, however, is clouded by the problems of determining the fit of the data to the model, and substantial additional research is necessary on this issue before questions of the invariance of the model can be adequately investigated.

Other developments. Because the IP model has frequently been used with existing M-C tests, Keats (1974) developed a IPL model with guessing. White (1976) derives this model from a CTT approach, but Colonius (1977) indicates that Keats' model results in no consistent maximum likelihood estimate for its parameters.

One advantage of IRT models is their ability to generalize beyond the binary responses commonly obtained from M-C tests to take into account information in incorrect responses to test items. Anderson (1977) and Andrich (1978a, 1978b) discussed generalizations of the IPL model to polychotomous items, which result in a successive integers scoring technique. Douglas (1978) develops estimation procedures for Andrich's model. One important characteristic of these models is that like the dichotomous case of the IPL model, integer scoring using equally distant weights preserves the IPL model characteristics. As a consequence, complex scoring procedures, such as are characteristic of the other IRT models are not required.

#### 2- and 3-Parameter IRT Models

The 2-parameter (2P) and 3-parameter (3P) IRT models are simply generalizations of the IPL model, including additional parameters that describe aspects of the IRF. The 2P model permits items to vary in the discrimination parameter, and the 3P model adds the lower asymptote (pseudo-chance value) to the IRF. Being generalizations of the IP model, the applications and utility of these models are essentially the same. That is, they have the capability of providing

sample-free measures of individuals, resulting in the same degree of "objectivity" as does the IPL model. These IRT models also permit the measurement of individuals with any subset of items, although the number-correct score for these models does not convey the same information as it does for the IPL model. Consequently, new scoring methods have been developed to implement these models, as have additional methods for the estimation of item parameters (Urry 1976; Wood, Wingersky, & Lord 1976).

One major advantage of IRT models is that the concept of reliability is not emphasized. The consequence is that all the confusion that has been engendered with regard to this concept in classical test theory disappears; and issues of homogeneity, internal consistency, type of reliability coefficient, and lower bounds are eliminated. In place of reliability, IRT uses the concept of information or precision of measurement, which is related to the standard error of measurement (or estimate) for a given level of a trait. Consequently, IRT permits the error of measurement to vary as a function of the variable being measured, and information and its derivatives (the conditional standard error of measurement or estimate) index this change in precision of measurement as a function of the trait being measured.

Samejima (1977a, 1977b) differentiates various aspects of the information function and provides critiques of the concept of reliability. She also develops the concept of weakly parallel tests (Samejima, 1977a), which are tests that have similar information functions but do not require the number of items, score categories, or other aspects of the tests to be similar. This redefinition of parallel tests permits not only the easier design of parallel tests for applied purposes but the conceptual definition of parallel adaptive/tailored tests. Samejima also provides criticisms of the classical standard error of measurement, indicating its group dependency (whereas the standard errors of measurement of IRT are not group dependent) and its dependence upon the heterogeneity of the group with regard to the trait being measured.

Parameter estimation and equating. As with the 1P IRT models, an important problem is the development of accurate methods for estimating the parameters of test items. This is somewhat more complex in the 2P and 3P models, since the problem becomes one of simultaneously estimating two or three parameters for each item plus an ability (trait) parameter for each person in the item calibration sample. Jensen (1976) proposes a direct conversion method for estimating IRT parameters from the item parameters of CTT. Schmidt (1977) evaluates a graphical method of direct conversion proposed earlier by Urry (1976). Ree (1979) compares four methods of estimating IRF parameters and concludes that no one of the procedures was consistently best, since the results obtained depended upon the characteristics of the data, while Waller (1980) studied yet another item parameterization approach under conditions of nonsymmetric distributions of ability. Samejima (1977a) describes a method of estimating the parameters of IRFs when previous estimates of ability are available for a group of individuals.

Similar to the 1P model, there have been several studies of the robustness of the 2P and 3P models under a variety of conditions. Ree and Jensen (1980) studied the effects of errors in item parameters on linear equating while Hambleton and Cook (1980) studied the robustness of the models under a variety of conditions, as well as the effects of test length and sample size on estimates

of the precision of latent trait ability scores. Reckase (1979) addressed his attention to the effects of multidimensionality in an item pool on item parameter estimates obtained in the 1P versus 3P models, and Lord (1975c) solved an empirical problem of the correlation between difficulty and discrimination parameters by redefining the ability scale onto a different metric. All these studies assist in gaining a better understanding of the potential of IRT models to perform adequately under a variety of situations.

The 3P models have also been applied to the problem of score equating and linking of items into larger pools. Marco, Petersen, and Stewart (1980) examined the adequacy of IRT score equating models when sample and test characteristics are systematically varied; Yen (1980) studied the effects of context on item parameter and trait estimates; and Ree and Jensen (1980) studied the effects of errors in item parameter estimates on linear equating.

Applications: Option weighting, adaptive/tailored testing. One of the problems that has not adequately been resolved by CTT is the problem of extracting additional information from the responses of testees to the incorrect options on a M-C item. Thissen (1976) addressed this problem directly using a polychotomous IRT model and found that it gave one-third to one-half more information than did the dichotomous model applied to the same data; an interesting subsidiary finding was that the reliability of the two models did not differ substantially, indicating the ineffectiveness of reliability as an index of the utility of different approaches to scoring items. Samejima (1977c) described another application of polychotomous latent trait IRT approaches. Bejar (1977) applied the continuous IRT model to personality assessment and found a good fit of the model to some of the personality data. His results, also evaluated in terms of information or precision of measurement, show considerable gains by use of this model over the usual dichotomous model.

Adaptive testing is the interactive administration of tests such that items are selected dynamically for each individual contingent upon the individual's responses to previous test items. Adaptive testing requires immediate scoring of each response and some means of selecting the next item to be administered on the basis of response information and/or ability estimates determined for each individual on an item-by-item basis. Although adaptive testing does not require IRT (Brooks & Hartz 1978; Hornke & Sauter 1980; Vale & Weiss 1975; Waters 1977; Weiss 1974), IRT has facilitated the development and implementation of most adaptive testing strategies. The review period has seen considerable progress on adaptive testing and the implementation of the 2P and 3P IRT models. Two major conferences have provided a forum for the discussion of current research in this field (Weiss 1978, 1980), while others (Jensema 1977; Lord 1977a; McBride 1977; Urry 1977) have pursued basic and applied research on the development and evaluation of a variety of adaptive testing strategies. These studies show, in general, that IRT combined with adaptive testing techniques is a viable methodology for the improvement of tests of ability and achievement and has considerable promise for the replacement of paper-and-pencil tests with computer-administered adaptive tests in the foreseeable future.

As might be expected, a few studies have been concerned with comparisons of IRT and CTT approaches. Douglas, Kahalil, and Farber (1979) compared CTT and IRT item analysis procedures by selecting items using traditional proportion-correct and item-total biserial correlations versus item selection based on 1PL

procedures. Their data show that about half the items were selected in common by the two procedures, some were selected by neither, and some by either. The correlation of proportion correct with IPL difficulty was .997, while IPL ability estimates correlated .91 with total score and .81 with score on the items selected by the CTT item selection procedure. In terms of validity, neither correlated differently with the criterion score. Their conclusion was that the two procedures define "different constructs," but there was no data to indicate which more adequately defined the trait desired. Lord (1977b) compared an IRT approach with three other approaches in the evaluation of the optimal number of choices in a test item. His results show that decreasing the number of choices per item, while lengthening the test proportionately, decreases the efficiency for low ability testees and increases the efficiency for high ability testees; his data also show that reliability comparisons of the methods do not demonstrate differences, whereas comparisons in terms of information (efficiency) describe differences in the characteristics of the different items.

Relationships with other psychometric models. One of the potentially most valuable contributions of IRT to psychological measurement is reflected in a series of papers relating the logic and procedures of IRT models to the mainstream of psychological measurement. A major deficiency of CTT has been in the separation of its logic and methodology from the other methods of psychological measurement. The methods of CTT are unique to that approach and have never been demonstrated to derive from or relate to any other models of psychological measurement. However, recent and important research during the review period has defined and described the continuity of the logic of IRT approaches with a variety of other approaches to psychological measurement.

That IRT approaches are a special case of Thurstone's scaling techniques is well demonstrated by Lumsden (1980), Brogden (1977), and Andrich (1978d). Wainer, Fairbank, and Hough (1978) analyzed a data set by both Thurstone scaling methods and the IPL model and demonstrate the similarity of the results. Perline, Wright, and Wainer (1979) and Brogden (1977) describe relationships between IRT models and additive conjoint measurement. Finally, an IRT model that implements the standard Likert successive integers attitude scaling approach has been developed (Andersen 1977; Andrich 1978a, 1978b, 1978c; Douglas 1978). Thus, by the use of IRT models, researchers can be assured of some continuity between test theory and other areas of psychological scaling.

Person fit. A major advantage of IRT models is the possibility of determining whether a person (or item) is performing in accordance with the assumptions of the models. Since the models make strong assumptions about the behavior of individuals and items, it is necessary to determine whether both individuals and items fit a given version of the model in order to adequately use it. If a set of individuals and items can be determined to be operating in accordance with the model, strong inferences can be made on the basis of the data and all of the power of the models can be put to practical use. If the responses of an individual (or a set of individuals to an item) do not fit the model, it can be concluded that the model is an inappropriate means of describing the behavior of that individual on that set of items (or that item on that set of individuals); this kind of statement can be also translated into a matter of degree of person (or item) fit, which can potentially lead to indices of precision of measurement for a given individual. Indeed, IRT permits the statement of the error of measurement associated with a unique set of responses of an individual to a

set of test items. These data can also be used to study the fit of individuals to a set of test items and, hence, to the assumptions underlying IRT.

Most of the work on person fit has been done with the 1P model. This work is well described by Wright (1977b), Wright and Stone (1980), and Wainer and Wright (1980). The approach generally used in the 1P model involves a chi-square test fit of a persons by items response matrix to the predicted probabilities from the 1P model. Lumsden (1977, 1978) generalizes the issue to one of person reliability. He defines the person characteristic curve (PCC), which has relationships to the observed data values used in the 1P chi-square index of person fit, and describes how the IRF and group reliability are functions of a series of PCCs. The idea of the PCC is redefined by Trabin and Weiss (1979) as the person response curve (PRC) to emphasize that it results from the responses of one individual to a set of test items. The PRC is traced back to work in the 1940s by Mosier, and some of the implications of it for the measurement of person fit are described. Trabin and Weiss derive the PRCs for a group of testees and proceed to test the fit of those testees to the 3P model. The results indicate an overwhelming fit of these individuals to the model, with the identification of a few individuals who appear to have systematic lack of fit for various reasons. Levine and Drasgow (1980) and Levine and Rubin (1979) call the problem one of measuring "appropriateness" of M-C test scores. They define a series of appropriateness (person fit) indices and study the application of these indices via monte carlo simulations, in addition to real data. Their data illustrate the potential of some of their indices to identify lack of fit of individuals to IRT models.

Thus, this new area of research, which has developed during the review period, promises to be an especially important one for future applications, since it will permit the identification of individuals for whom IRT does not adequately describe their behavior in a testing situation. The result will be statements of individual precision for the test score of one person on a set of items, possibly resulting in an important moderator variable to be used in prediction studies to improve predictive validity.

#### Order Models

Another new area of research that has appeared during the last five years is the application of order-based models to the development of psychological measuring instruments. To differentiate them from ordinal test theory models, these models are based on the logical relationships among item responses (and individuals) utilizing items by persons dominance matrices. The methodologies have relationships with mathematical information theory (Krus & Ceurvorst 1979) and have their basic psychometric roots in earlier work by Guttman and in scalogram analysis (Airasian, Madaus, & Woods 1975; Bart 1976).

The majority of the research in order analysis has been in the field of attitude scaling in the analysis of the structure of item/person matrices (Bart 1978; Krus 1977, 1978; Krus & Weiss 1976) and in the analysis of instructional hierarchies (Airasian & Bart 1975; Bart & Mertens 1979). Cliff (1979), however, has translated the approach into a test theory approach that does not assume true scores. It is interesting to note that this approach also permits expressions of person consistency similar to the person fit approaches in IRT. Cliff then generalizes the application of his order theory methods to adaptive testing

(Cliff 1975, 1977; Cliff, Cudeck & McCormick 1979; Cudek, McCormick & Cliff 1979), whereas Baker and Hubert (1977) propose some inference procedures and hypothesis testing procedures for order theory. Initial results of order theory seem promising, but additional research in the test theory area is necessary to determine the degree of sample specificity of this approach if it is to provide any advantages over CTT. Since both order methods and IRT methods have their ancestry in Guttman's (1944) scalogram analysis, some thought should also be given to the relationships between the two methodologies.

#### Miscellaneous Models

A few additional new developments appeared during this period as alternatives to CTT. Wilcox (1979e) and Morrison and Brockway (1979) discuss applications of the beta-binomial model to testing problems. Mellenbergh, Kelderman, Stijlen, and Zondag (1979) develop linear models for the analysis and construction of measuring instruments using a facet (factorial) design, a special application of covariance structure analysis to the construction and analysis of measuring instruments. Their approach is an alternative to generalizability theory (and CTT) and permits design of instruments to fit a hypothesized facet-type structure. McQuitty (1976) describes an item analysis procedure based on configural approaches, while Schulman (1976, 1978; Schulman & Haden 1975) develops a test theory for ordinal measurements which arrives at the same kinds of definitions of reliability, attenuation, and errors of measurement as does CTT. It differs from order theory approaches in that it is basically an ordinal theory based on total scores as compared to the order theory approaches that are based on logical relationships among persons and items at the item level. Finally, Whitely and Dawis (1976) present and apply a model designed to psychometrically distinguish the concept of aptitude (potential) from ability (current status). Their data suggest that the predictability from later stages can be improved by adding the gains resulting from specific interventions. All these models attempt to generalize or to replace the deficiencies in the CTT model. All, however, will require additional research and development work before they become useful.

### VALIDITY

#### Content and Construct Validity

Two seemingly unrelated phenomena--the test fairness controversy (see below) and the CRT movement--have heightened interest in content validity (Schoenfeldt, Schoenfeldt, Acker, & Perlson 1976). Some believe that a content valid employment test or success criterion is inherently fair. Much of the literature on CRT has emphasized the content validity of educational achievement tests to the exclusion of construct and criterion-related validity. The heightened interest in content validity has led to a controversy about when or whether any test can be judged solely on the basis of content validity.

Ebel (1975) argues that construct validity is not a concern if the behavior can be directly observed or the trait can be operationally defined. In opposition to the increased emphasis on content validity in educational testing, Messick (1975) argues that construct validity is as important for educational tests as for psychological tests. In what could be considered a response to Ebel

(1975), he points out the logical difficulties associated with operational definitions. Guion (1977, 1978) presents his reservations about the increased emphasis on content validity in employment testing, including his concern that expert judgments about content validity are often made too glibly. He goes on to list six conditions which, in his opinion, a test must meet before it can be judged solely on the basis of its content validity, conditions which are much more stringent than those of Ebel (1975). Guion (1974) discusses the merits and limitations of all three kinds of validity; construct, content, and criterion-related validity. Several authors have considered the context of educational testing rather than the content validity of single tests: Carver (1974, 1975), Hoepfner (1974), and Levine (1976) lament the fact that published educational tests tap so narrow a set of educational objectives.

#### Multitrait-Multimethod Matrices

Structural equation models have been applied to the study of multitrait-multimethod (MTMM) correlation matrices in the search for statistical procedures useful in studying aspects of construct validity. Ray and Heeler (1975) compare restricted maximum likelihood factor analysis and multidimensional scaling as methods of analyzing MTMM matrices. According to Kalleberg and Kluegel (1975), the structural equations approach has the advantage that (1) it allows estimation of correlations between trait and method factors, (2) it provides estimates of both trait and method factor influences on each measure, and (3) it forces researchers to specify their assumptions. Mellenbergh et al. (1979) note that structural equations models can be extended to the study of any test facet model, of which the MTMM model is one example and Guilford's (1967) structure of intellect model is another.

Avison (1978) and Schmitt (1978) point out that there is not just one but several structural equations models for studying MTMM matrices. Schmitt (1978) discusses the problem of choosing between possible models on the basis of their fit to the data, a problem that is only partially solved at present. It is not clear whether the choice of model substantially influences the conclusions reached.

Methods of investigating MTMM matrices that do not rest on structural equations models have been described by Golding (1977), Golding and Seidman (1974), Hubert and Baker (1978, 1979), Jackson (1975, 1977), Levin (1974), and Lomax and Algina (1979). After reviewing alternatives to the structural equations approach, Schmitt, Coyle, and Saari (1977) conclude that the structural equations models provide the most detailed information about individual traits and methods.

The structural equations model for MTMM data contains an implicit definition of method variance, a term which Campbell and Fiske (1959) left only vaguely specified (Golding 1977). That is, method variance is variance attributable to a dimension of individual differences that (1) is uniquely associated in the factor pattern matrix with measures employing one particular method of measurement, (2) contributes to the variance of any measure assessed by that method, and (3) combines in an additive fashion with other sources of variance. Other definitions are possible. In Tucker's (1966) three-mode factor model for MTMM matrices, for instance, traits and methods combine in a multiplicative interaction rather than in an additive fashion. Because the structural equations ap-

proaches are becoming widely accepted, the definitions of trait and method variance implicit in those models deserve closer scrutiny than they have received in the past. Tesser and Krauss (1976) remind us that a MTMM is not the only way to investigate construct validity.

### Predictive Validity

How large a sample size is needed to study a test's predictive validity? This is the question addressed by Cascio, Valenzie, and Silbey (1978) and Schmidt and Hunter (1977). Schmidt and Hunter argue that the sample sizes needed for predictive validity studies are often much larger than commonly recommended. Because the observed correlation is typically reduced by such influences as restriction in range and criterion unreliability, large sample sizes are needed to insure adequate power in statistical tests of predictive validity coefficients.

Schmidt and Hunter (1977) and Schmidt, Hunter, Pearlman, and Shane (1979) argue against the dominant belief that the predictive validity of selection tests is highly situation specific. Prior research has revealed considerable variation in the observed validity coefficients for the same test in several job settings. Schmidt and his coworkers argue that most of this variation is not due to fluctuations in the true validity of the test. Rather, much of the variation is due to artifactual sources, including variation from one job setting to the next in (1) criterion reliability, (2) test reliability, (3) range restriction, and (4) criterion contamination. Because of the small sample sizes used in many validity studies, sampling error can also account for some of the variation. Schmidt and Hunter (1977) propose a Bayesian method of combining validity coefficients across studies on the same job family to arrive at pooled estimates of validity.

As Schmidt and his coworkers argue, a portion of the variation in a test's validity coefficient from study to study is due to artifactual sources and sampling error. How much is due to those sources? Schmidt and his coworkers pile one untested assumption upon another to arrive at their estimates and to develop their Bayesian approach. The Bayesian alternative is only as good as the untested assumptions; and it presumes a satisfactory method of classifying tests into job families, something which does not now exist. Callender and Osburn (1979) and Callender, Osburn, and Greener (1979) propose an alternative model that leads to smaller estimates of the artifactual variance and to an alternative Bayesian approach. Rock, Werts, Linn, and Jöreskog (1977) provide some possible methodological assistance for the criterion-related validity problem in their structural equations model that partitions criterion variance into (1) measurement error, (2) true score variance accounted for by the predictor, and (3) true score variance unaccounted for by the predictor.

Although Schmidt and Hunter's (1977) Bayesian model may not be the answer, their work raises an important issue. Given the often unavoidable limitations (particularly limitations of sample size) in job specific validity studies, would pooled estimates sometimes be better? If so, under what conditions, and how should the several job specific coefficients be pooled? A workable taxonomy of job families would need to be developed before job pooling could become accepted (Pearlman 1980).

Schulman (1976) presents a predictive validity model for ordinal measurements. A surprising number of authors have examined the validity of self-report ability measures (DeNisi & Shaw 1977; Farrel 1979; Levine & Flory 1977; Norris 1976; Norris & Chapman 1976; Pohlmann & Beggs 1974). Hogan, De Soto, and Solono (1977), Mischel (1977), and Wade and Baker (1977) ponder the value of personality tests. The economic impact of valid selection was examined by Schmidt, Hunter, McKenzie, and Muldrow (1979).

#### Moderator and Suppressor Effects

Lissitz and Schoenfeldt (1974), Gross, Steckler-Faggen, and McCarthy (1974), and Novick and Jackson (1974) consider the problem of using subgroup information as a moderator variable in prediction equations. Drösler (1978) presents a scheme for increasing the temporal range of psychometric predictions. Conger (1974) and Velicer (1978) attempt to improve the definition of suppressor variables and methods for dealing with them, whereas McFatter (1979) illustrates a structural equations approach for interpreting suppressor and enhancer variables. Brown (1979), Greener and Osburn (1979), Gullickson and Hopkins (1976), and Roe (1979) consider the accuracy of corrections for restriction in range. Sands and Alf (1978) present a correction for restriction in range that does not require that the user know the variance of the predictor in the applicant population, although it does require knowledge of the selection ratio. Osburn and Greener (1978) discuss methods of sampling selected applicants for inclusion in a predictive validity study when criterion information is too difficult or expensive to collect from all selected applicants.

#### Educational Applications

Cronbach and Snow's (1977) book on aptitude and instructional methods is a landmark review of the research on the interaction between instructional methods and student aptitudes. The authors conclude that the literature contains very few examples of consistently replicated interactions between measured aptitudes and instructional methods. Hunt (1975) suggests that new types of tests will need to be developed--tests that are specifically designed to assess those characteristics that interact with educational methods. Corno (1979), Tobias (1976), and Winne (1977) report finding isolated aptitude-treatment interactions of various types, all of which require further replication. It is clear from Cronbach and Snow's (1977) review that research will only slowly reveal how person characteristics and instructional methods interact. An understanding of such interactions would greatly enhance the ability to adapt instruction to the learner's needs.

Airasian and Bart (1975), Dayton and MacReady (1976), Davison (1980), and Davison and Thoma (1980) describe methods for studying the internal structure of tests constructed around hypothesized item hierarchies. Davison (1977, 1979) has discussed methods of studying the interrelationships between subscales, each of which corresponds to an ordered stage in a developmental sequence. Applications of these techniques can be found in Davison, King, Kitchener, and Parker (1980), Davison and Robbins (1978), Davison, Robbins, and Swanson (1978), and Jepsen and Grove (1980).

### Test Fairness

On many educational and occupational selection tests, some American minorities--Blacks, Hispanics, Native Americans, and some Asian Americans--form populations with lower mean scores than the White majority. Not all ethnic minorities, however, have consistently lower mean scores, notably Chinese and Japanese Americans. For those minorities with lower mean scores, the result can be a lower rate of selection for jobs or educational admission if selection is based heavily on tests. However, there is little information available about how heavily test information influences selection decisions. Without such information, it is impossible to say how much of a barrier tests have actually posed to minorities seeking selection to jobs or admission to educational institutions.

Three books on test fairness have appeared during the period of this review. Neither Oaklund's (1977) nor Samuda's (1975) books scrutinize alternatives to standardized tests with the same critical eye with which they evaluate traditional tests. Nor do they present the case in favor of standardized testing with the same thoroughness with which they present the case against. Jensen's (1980) defense of test fairness is a more up-to-date and complete treatment. His thoroughness is attested to by the fact that the popular press seems to draw from his work even to criticize testing (for instance, compare Jensen 1980, p. 5, with Sewall, Carey, Simons, and Lord 1980, p. 97).

Fairness of tests to women has also been of concern. The context of this discussion is quite different, however, because the mean scores of females on many tests, particularly verbal aptitude tests, is higher than that of men. Where women do have lower average scores, the differences are often not as marked as for racial minorities. Maccoby and Jacklin's (1974) work on sex differences aids in understanding the discussion of test fairness to women.

### Definitions of Test Fairness

Various authors have proposed definitions of bias in selection, bias in a test, and bias in a test item.

Bias in selection. There are at least five major definitions of bias in selection. In general, no selection strategy can satisfy all of the fairness definitions. According to Cleary (1968, p. 115), a test is biased against members of a subgroup "if in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup." Einhorn and Bass (1971) define selection as fair if the least qualified persons who would be accepted from each subgroup have an equal chance of succeeding. Several authors define fairness in terms of ratios. Selection can be defined as fair if the ratio of the number selected to the number qualified is the same for all subgroups (Thorndike 1971), if the ratio of the number selected and qualified to the number qualified is the same for all groups (Cole 1973), or if the ratio of the number selected and qualified to the number selected is the same for all groups (Linn 1973). Definitions of fairness are critiqued in Bernal (1975), Cleary, Humphrey, Kendrick, and Wesman (1975), Cronbach (1976), Darlington (1976, 1978), Flaughner (1978), Hunter and Schmidt (1974, 1976, 1978a), McNemar (1975), Myers (1975), Novick and Ellis (1977), Peterson and Novick (1976), Pine and Weiss (1976), and Sawyer, Cole, and Cole (1976).

Peterson and Novick (1976) point out serious logical inconsistencies in the three ratio models. Hunter, Schmidt, and Rauschenberger (1977) note that the Cleary (1968) model seems to be the only one adopted by the U.S. Equal Employment Opportunity Commission (1970) guidelines on employee selection, and it has been upheld in a U.S. District Court decision (Cortez v. Rosen, Northern District of California, March 11, 1975). If the courts tightly limit the use of race in selection, as Novick and Ellis (1977) suggest, then use of the above fairness models would be correspondingly limited. Recent U.S. Supreme Court decisions, the so-called Bakke and Weber decisions, however, suggest that racial information can be used even by institutions that have no prior history of discrimination.

A decision theoretic approach in which institutions assign utilities to selection outcomes, utilities that may vary as a function of the race or sex of the applicant, have been endorsed by Darlington (1976, 1978), Gross and Su (1975), Linn (1976a), Petersen and Novick (1976), and Sawyer, Cole, and Cole (1976). As a criterion for selection fairness, decision theory is hopelessly vague. As a consequence, it can be used to discriminate against any desired group by appropriately assigning utilities to outcomes. Petersen and Novick (1976), however, show that a decision theoretic framework can profitably be used to evaluate various proposed fairness models. Tools used to equalize the proportion of majority and minority members selected now include quotas (Rybak 1980), bonus points for minority or disadvantaged applicants (Roark 1978), and separate standardization of a test by subgroups so that the test has the same mean in minority and majority subgroups (Mercer & Lewis 1978).

Not all definitions of bias describe bias in selection. Jackson (1975) presumes that Blacks and Whites are equal in ability and, therefore, that any test is biased if the mean scores for Blacks and Whites are different. Faggen-Steckler, McCarthy, and Tittle (1974) propose a measure of item content bias. Removing bias in test content, however, need not affect differences between mean test scores of groups. Echternacht (1974), Ironson and Subkoviak (1979), and Scheuneman (1979) discuss performance-based measures of item bias. By eliminating items that contain bias as assessed by one of these performance-based measures, the most biased items in the test may be eliminated. After pruning such items, however, the test itself will be unbiased only if the average item in the original item pool was unbiased (Green 1978). Flaughner and Schrader (1978) found that pruning biased items did not substantially alter the mean difference between minority and majority students and, hence, that such methods of pruning items would likely not materially affect the adverse impact of selection decisions.

#### Fairness to Minorities

Empirical studies of tests administered to racial or ethnic minorities have focused heavily on Blacks and to a lesser extent on Mexican Americans. There was much less research on Native Americans, Asian Americans, and Non-Mexican Hispanic populations. The most thoroughly researched setting was the college admissions situation in which the predictors are high-school grade-point average (GPA) and scholastic admissions tests and in which the criterion is college GPA. Although there were numerous studies of employment selection, there was little consistency in the predictors and criteria employed. There are some general trends in the employment studies, but no conclusions can be drawn about specific

jobs, tests, or criterion variables. As has been noted time and time again by authors in the area, the research strategies assume an unbiased criterion, when in practice there is no agreed upon standard by which to judge the criterion. Conclusions to be drawn from the research described below depends upon whether the criteria employed are believed to be biased or not biased.

Differential and single-groups validity. It is commonly stated that traditional tests are less valid for minority applicants than for nonminorities. Such statements have given rise to the single-group and differential validity issues. Differential validity is said to exist when predictive validity coefficients are unequal in the minority and majority subgroups ( $\rho_1 < \rho_2$ ). When the predictive validity coefficient is greater than zero for only one subgroup ( $0 = \rho_1 < \rho_2$ ), then single group validity is said to exist. In her seminal work, Boehm (1972) defined single group and differential validity differently, but her definitions contain logical problems (Bartlett, Bobko, & Pine 1977; Hunter & Schmidt 1978b).

The research on differential and single-group validity has been plagued by methodological problems including nonindependence of observations, statistical tests with low power, and differential restriction of range in majority and minority populations. Bartlett, Bobko, and Pine (1977), Bobko and Bartlett (1978), Boehm (1977, 1978), Hunter and Schmidt (1978b), Hunter, Schmidt, and Hunter (1979), Katzell and Dyer (1977, 1978), and O'Connor, Wexley, and Alexander (1975) have discussed these methodological issues. Conclusions about the differential validity hypothesis have varied, depending on whether or not the reviewer included data on tests whose validity did not differ significantly from zero in any sample and which, hence, would not be used as a selection device.

In the area of employment selection, Bobko and Bartlett (1978), Boehm (1977, 1978), Gael, Grant, and Ritchie (1975a, 1975b), Hunter and Schmidt (1978b), Hunter, Schmidt, and Hunter (1979), Linn (1978b), O'Connor, Wexley, and Alexander (1975), and Reeb (1976) present and interpret the evidence pertaining to single group and differential validity. Particularly in later studies (Boehm 1977; Hunter, Schmidt, & Hunter 1979; Katzell & Dyer 1977; Linn 1978b; O'Connor et al. 1975), authors conclude that the evidence against the single group validity hypothesis is overwhelming. Authors still differ on whether or not examples of differential validity occur more frequently than can be attributed to artifacts and sampling error. Most reviewers, however, conclude that examples of differential validity are rare and that when differences in validity do exist, they are usually small. Boehm (1977) found that the most methodologically sound studies reported the fewest examples of differential and single group validity. Both Bobko and Bartlett (1978) and Linn (1978b) conclude that the single and differential group validity issues are secondary to the question of whether or not the performance of minorities is systematically underpredicted by tests. We strongly agree.

In the educational literature, examples of large differences in minority and majority validities are just as rare as in the employment literature. Wright and Bean (1974) found that the college GPAs of high socioeconomic status (SES) students were somewhat better predicted than those of low SES students. Pfeifer (1976) found little difference in the predictability of Whites and Blacks. Breland (1978) and Wilson (1978) concluded that the traditional predictors of college GPA are generally valid predictors for both majority and minor-

ity students. Flaughner (1978) argues that if educational examples of single group and differential validity are so difficult to find, then they are probably not of much practical import.

Comparisons of regression lines. Although some researchers have been studying differential validity, others have been studying test fairness as defined by Cleary (1968). That is, they have been examining regression lines to determine if use of a common regression line for both majority and minority subgroups would result in over- or under-prediction of success for either group. Goldman and his coworkers (Goldman & Hewitt 1976, 1975; Goldman & Richards 1974; Goldman & Widawski 1976), generally found no evidence for bias in the prediction of college grades among Blacks, Whites, Chicanos, and Orientals in the University of California system. In one exception (Goldman & Hewitt 1975), the authors found trivial differences between regression lines for Anglo and Mexican American samples. In another series of studies at California colleges, Warren (1976) found only two instances in which regression lines were significantly different for Anglo and Mexican Americans. In one case, selection was biased in favor of Mexican Americans; in one case it was biased against them; and in both cases the bias was small. Cleary et al. (1975) reviewed several studies comparing regression lines for Blacks and Whites, concluding that when only standard courses are included in the college GPAs, differences in the regression lines are small and favor Blacks more often than Whites. Silverman, Barton, and Lyon (1976) found bias in favor of Blacks. When differences exist, it is usually because the regression lines for the two groups have different intercepts rather than because they have different slopes.

What can be concluded from these studies based on Cleary's (1968) definition of fairness? In the published literature, the evidence suggests that tests do not consistently underpredict the performance of minorities on traditional success criteria when a common regression line is used for both the majority and minority groups. This means that the tests are no more or less biased than the criteria they are designed to predict. The evidence could be said to overwhelmingly support the fairness of tests were it not for lingering doubts about the fairness of traditional success criteria. There is a pressing need to define what constitutes a fair criterion and then to evaluate traditional success criteria against that definition. Without further work on the criterion problem, a more definitive answer to the question of test fairness is impossible within the Cleary framework.

Adverse impact. A somewhat different approach to the study of test fairness was adopted by Hunter, Schmidt, and Rauschenberger (1977). They compared the adverse impact and validity of selections based on four fairness strategies. Cleary's (1968) model was the most valid and a quota model had the least adverse impact. The most valid models also had the greatest adverse impact and vice versa. Breland and Ironson (1976) used admissions data from the University of Washington Law School to compare the Cleary (1968), Cole (1973), and Thorndike (1971) definitions of fairness in terms of the number of minority applicants who would be selected, using selection rules satisfying each. Differences between fairness models were small. Because none of the fairness models would have selected as many minority applicants as did the admissions committee, Breland and Ironson (1976) argue against the adoption of any psychometric fairness model, and for the values embodied in selection committee decisions. This is a curious argument in light of the historical fact that the search for a fairness defini-

tion began as a reaction to seemingly unfair personnel and admissions decisions, particularly in the South, and that the object of the search was to find a fairness standard by which admissions and selection decisions could be judged.

Bias in test content. It is commonly suspected that test content is biased. Smeiser and Ferguson (1978) found that mean scores of Whites were higher than mean scores of Blacks, even when the test material was written from a Black perspective. In the Smeiser and Ferguson (1978) study, the cultural information needed to answer the items was provided in reading passages and did not need to be recalled by the examinee. Successfully answering the items depended only on correct reasoning with the information given. It is, however, possible to reduce or even to reverse the mean difference between majority and minority groups using items for which successful completion requires recall of information more commonly available in the minority subculture (Medley & Quirk 1974; Williams 1975). No one appears to have investigated the predictive utility of tests with content constructed so as to reduce subgroup differences. A traditional test would presumably be better than such a nontraditional test for predicting traditional academic or employment outcome criteria in a racially heterogeneous population, because the traditional test could better account for individual differences in criterion performance between members of the different subpopulations. Jensen (1974) found that race of the examiner seldom affected mean scores of examinees. In the one exception, both Whites and Blacks had higher scores in the presence of a White examiner.

Alternatives to tests. A number of papers discuss non-paper-and-pencil alternatives to tests, primarily subjective evaluations by personnel officers, employment supervisors, or teachers (i.e. grades). Arvey (1979), Hamner, Kim, Baird, and Bigness (1974) discuss the problems of bias in employer evaluations. Cascio (1976) found that biographical items were equally valid for majority and minority applicants. Goldman and Widawski (1976) point out that minority and majority mean differences are typically smaller on high school grades than on standardized tests. They suggest that in some settings, selecting among college applicants solely on the basis of high school grades would increase the number of minorities selected without materially affecting the validity of the selections.

#### Fairness to Women

Reed (1976) reviewed the sex differences literature on test fairness as defined by Cleary. She concluded that traditional predictors often underpredict the performance of college achievement for females. She points out that further investigation is needed into the reasons why. For example, the differences in regression lines may result because (1) males and females typically enter different fields of study or (2) females who enter college are a more select sample of the female population. If these two explanations are correct, differences in male and female regression lines would be expected to decrease as more women enter male-dominated college majors and a greater proportion of women enter college.

Reilly, Zedeck, and Tenopyr (1979) studied physical measures (e.g. arm strength, height) as predictors of performance in an outdoor craft. No differences in regression lines were found, suggesting that selection based on such criteria is fair, as defined by Cleary (1968). In studies of differential va-

lidity, Gross, Faggen, and McCarthy (1974) and Schmitt, Melton, and Bylenga (1978) found traditional measures consistently predicted academic criteria slightly better for females than for males. Schmitt et al. (1978) found the same trend for employment data, but the number of employment studies reviewed was small. Moss and Brown (1979) found that varying the sex referant in reading passages did not significantly change the reading comprehension scores of males and females. Since the information needed to successfully answer the questions seems to have been given in the reading passage, the result is not surprising.

Mai-Dalton, Feldman-Summers, and Mitchell (1979), Simas and McCarrey (1979), Arvey (1979), Hamner, Kim, Baird, and Bigoness (1974) have examined the fairness to women of employee interviews and job performance evaluations. The direction and degree of sex bias in these studies appears to be a complex interaction of the rater sex, ratee sex, and job characteristics. Arvey (1979) notes that employment interviews may be discriminatory if women are asked different questions (e.g. What will you and your husband do when your children get sick?).

#### SUMMARY AND CONCLUSIONS

The period 1975 through 1979 has had considerable activity in test theory and its methods, covering a diverse range of topics. Because the main results and methods of classical test theory were developed and refined over the last 70 years or so, little progress was made in classical test theory, since there is little progress to be made. The period saw active work in developing alternatives to classical test theory. Item response theory, particularly its applications to a variety of testing problems inadequately handled by classical test theory, has been the subject of considerable research activity. Methods and procedures for both the Rasch model and the generalizations of the Rasch model to more complex item response functions have been the objects of a considerable amount of research. Estimation procedures for these models have been refined and investigated, and the robustness of the estimation procedures has been studied under a variety of circumstances. The result is the beginning of a better appreciation of the promise and limitations of these models and their areas of application. Progress has been made in the development of equating procedures using IRT models and in their applications to adaptive testing. An important new field of research that has developed as a result of the use of IRT models is the area of person fit (person reliability, or appropriateness measurement), which has considerable promise for applications of psychological measurement in practical situations. In addition, considerable research remains yet to be done on IRT models. The period has seen a needed integration between test theory approaches based on IRT and other models of psychological measurement. More work is needed in this area to specify and to describe the relationships of IRT models to other areas of psychological measurement, in order to reintegrate psychological testing into the mainstream of psychology and its measurement procedures.

There has been more research, and less speculation, about the utility of criterion-referenced tests during this period. Some technical advances have been made, but the problem of the arbitrariness of the cutting scores still remains a serious limitation to important applications of these methods. Order theory has developed as a possible viable approach to psychological testing, but

considerable additional research is needed before it can be shown to have definite advantages over that of either classical test theory or item response theory. No studies are yet available comparing order theory and item response theory approaches on the same data sets.

Issues of test fairness have received considerable attention. The literature has focused on problems of item and test bias and on test fairness in the study of differential validity. Before the issue of test fairness can be adequately resolved, the problem of fairness of criteria remains yet to be addressed. However, the search continues for selection devices other than tests that are likely to be less unfair. A realistic comparison of these approaches, however, would include evaluation of these alternatives on the same criteria used to evaluate the tests themselves.

Some progress was made in the area of validation by the use of structural equation models, particularly in the analysis of multitrait-multimethod matrices. The area of content validity was somewhat more adequately defined, but the issues still reduce to an unacceptable degree of individual judgment for the definition of content validity. Some research during the period has contributed to problems in the understanding of predictive validity.

Thus, similar to most other fields, progress comes slowly. Future research in test theory will make more progress if less emphasis is placed on relatively trivial research in classical test theory and on the derivation of new formulas for already known concepts, and more emphasis is placed on the evaluation of alternative models that promise considerable improvement in the design, construction, and implementation of psychological measuring instruments.

## REFERENCES

- Abu-Sayf, F. K. 1977. A new formula score. Educational and Psychological Measurement 37:853-862
- Abu-Sayf, F. K., & Diamond, J. C. 1976. Effect of confidence level in multiple-choice test answers on reliability and validity of scores. Journal of Educational Research 70:62-63
- Aiken, L. R. 1979. Relationships between item difficulty and discrimination indexes. Educational and Psychological Measurement 39:821-824
- Aiken, L. R., & Williams, E. N. 1978. Effects of instructions, option keying, and knowledge of test material on seven methods of scoring two-option items. Educational and Psychological Measurement 38:53-59
- Airasian, P. W., & Bart, W. M. 1975. Validating a priori instructional hierarchies. Journal of Educational Measurement 12:163-174
- Airasian, P. W., Madaus, G. F., & Woods, E. M. 1975. Scaling attitude items: A comparison of scalogram analysis and ordering theory. Educational and Psychological Measurement 35:809-819
- Algina, J. 1978. Comments on Bartko's "On various intraclass correlation reliability coefficients." Psychological Bulletin 85:135-138
- Algina, J., & Noe, M. J. 1978. A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement 15:101-10
- Allison, P. D. 1975. A simple proof of the Spearman-Brown formula for continuous length tests. Psychometrika 40:135-136
- Andersen, E. B. 1977. Sufficient statistics and latent trait models. Psychometrika 42:69-81
- Andersen, E. B., & Madsen, M. 1977. Estimating the parameters of the latent population distribution. Psychometrika 42:357-374
- Andrich, D. 1978a. A binomial latent trait model for the study of Likert-style attitude questionnaires. British Journal of Mathematical and Statistical Psychology 31:84-98
- Andrich, D. 1978b. Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement 2:581-594
- Andrich, D. 1978c. A rating formulation for ordered response categories. Psychometrika 43:561-573
- Andrich, D. 1978d. Relationships between the Thurstone and Rasch approaches to

- item scaling. Applied Psychological Measurement 2:451-462
- Arvey, R. D. 1979. Unfair discrimination in the employment interview: Legal and psychological aspects. Psychological Bulletin 86:736-765
- Avison, W. R. 1978. Auxiliary theory and multitrait-multimethod validation: A review of two approaches. Applied Psychological Measurement 2:433-449
- Baker, F. B., & Hubert, L. J. 1977. Inference procedures for ordering theory. Journal of Educational Statistics 2:217-233
- Bart, W. M. 1976. Some results of ordering theory for Guttman scaling. Educational and Psychological Measurement 36:141-148
- Bart, W. M. 1978. An empirical inquiry into the relationship between test factor structure and test hierarchical structure. Applied Psychological Measurement 2:333-337
- Bart, W. M., & Mertens, D. M. 1979. The hierarchical structure of formal operational tasks. Applied Psychological Measurement 3:343-350
- Bartlett, C. J., Bobko, P., & Pine, S. M. 1977. Single-group validity: Fallacy of the facts. Journal of Applied Psychology 62:155-157
- Bejar, I. I. 1977. An application of the continuous response level model to personality measurement. Applied Psychological Measurement 1:509-521
- Bejar, I. I., & Weiss, D. J. 1977. A comparison of empirical differential option weighting scoring procedures as a function of inter-item correlation. Educational and Psychological Measurement 37:335-340
- Bentler, P. M. 1980. Multivariate analysis. Annual Review of Psychology 31:419-456.
- Berk, R. A. 1978. Empirical evaluation of formulae for correction of item-to-total point-biserial correlations. Educational and Psychological Measurement 38:647-652
- Bernal, E. M., Jr. 1975. Comment on "Educational uses of tests with disadvantaged students." American Psychologist 30:93-95
- Beuchert, A. K., & Mendoza, J. L. 1979. A monte carlo comparison of ten item discrimination indices. Journal of Educational Measurement 16:109-117
- Bigoness, W. J. 1976. Effect of applicant's sex, race, and performance on employer's performance rating. Journal of Applied Psychology 61:80-85
- Block, J. H. 1978. Standards and criteria: A response. Journal of Educational Measurement 15:291-295
- Bobko, P., & Bartlett, C. J. 1978. Subgroup validities: Differential definitions and differential prediction. Journal of Applied Psychology 63:12-14

- Boehm, V. R. 1972. Negro-white differences in validity of employment and training selection procedures. Journal of Applied Psychology 56:33-39
- Boehm, V. R. 1977. Differential prediction: A methodological artifact? Journal of Applied Psychology 62:146-154
- Boehm, V. R. 1978. Populations, preselection, and practicalities. Journal of Applied Psychology 63:15-18
- Bond, L. 1979. On the base-free measure of change proposed by Tucker, Damarin, and Messick. Psychometrika 44:351-356
- Breland, H. M. 1978. Population validity and college entrance measures. Princeton, NJ: Educational Testing Service (ETS RB-78-19)
- Breland, H. M., & Ironson, G. H. 1976. Defunis reconsidered: A comparative analysis of alternative strategies. Journal of Educational Measurement 13:89-99
- Brennan, R. L. 1975. The calculation of reliability from a split-plot factorial design. Educational and Psychological Measurement 35:779-788
- Brennan, R. L. 1980a. Applications of generalizability theory. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press
- Brennan, R. L. 1980b. Handbook for Gapid: A Fortran IV computer program for generalizability analyses with single-facet designs. Iowa City: American College Testing Programs (Technical Bulletin 34)
- Brennan, R. L., & Kane, M. T. 1977. An index of dependability for mastery tests. Journal of Educational Measurement 14:277-289
- Brogden, H. 1977. The Rasch model, the law of comparative judgment and additive conjoint measurement. Psychometrika 42:631-634
- Brooks, S., & Hartz, M. A. 1978. Predictive ability of a branching test. Educational and Psychological Measurement 38:415-420
- Brown, J. M., & Weiss, D. J. 1976. An adaptive testing strategy for achievement test batteries. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (Research Report 77-6)
- Brown, S. H. 1979. Validity distortions associated with a test in use. Journal of Applied Psychology 64:460-462
- Burton, N. 1978. Societal standards. Journal of Educational Measurement 15:263-271
- Callender, J. C., & Osburn, H. G. 1977a. A method for maximizing split-half reliability coefficients. Educational and Psychological Measurement 37:819-825

- Callender, J. C., & Osburn, H. G. 1977b. An empirical comparison of coefficient alpha, Guttman's Lambda-2, and MSPLIT maximized split-half reliability estimates. Journal of Educational Measurement 16:89-99
- Callender, J. C., & Osburn, H. G. 1979. Development and test of a new model for validity generalization. (Manuscript submitted for publication)
- Callender, J. C., Osburn, H. G., & Greener, J. M. 1979. Small sample tests of two validity generalization models. (Paper presented at the annual convention of the American Psychological Association, New York)
- Campbell, D. T., & Fiske, D. W. 1959. Convergent and discriminant validity by the multitrait-multimethod matrix. Psychological Bulletin 56:81-105
- Cardinet, J., Tourneur, Y., & Allal, L. 1976. The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement 13:119-135
- Carver, R. P. 1974. Two dimensions of tests: Psychometric and edumetric. American Psychologist 29:512-518
- Carver, R. P. 1975. The Coleman Report: Using inappropriately designed achievement tests. American Educational Research Journal 12:77-86
- Cascio, W. F. 1976. Turnover, biographical data, and fair employment practices. Journal of Applied Psychology 61:576-580
- Cascio, W. F., & Kurtines, W. M. 1977. A practical method for identifying significant change scores. Educational and Psychological Measurement 37:889-895
- Cascio, W. F., Valenze, E. R., & Silbey, V. 1978. Validation and statistical power: Implications for applied research. Journal of Applied Psychology 63:589-595
- Claudy, J. G. 1978. Biserial weights: A new approach to test item option weighting. Applied Psychological Measurement 2:25-30
- Cleary, T. A. 1968. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement 5:115-124
- Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. 1975. Educational uses of tests with disadvantaged students. American Psychologist 30:15-41
- Cliff, N. 1975. Complete orders from incomplete data: Interactive ordering and tailored testing. Psychological Bulletin 82:289-302
- Cliff, N. 1977. A theory of consistency of ordering generalizable to tailored testing. Psychometrika 42:375-399
- Cliff, N. 1979. Test theory without true scores? Psychometrika 44:373-393

- Cliff, N., Cudek, R., & McCormick, D. J. 1979. Evaluation of implied orders as a basis for tailored testing with simulation data. Applied Psychological Measurement 3:495-514
- Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:37-46
- Cohen, L. 1979. Approximate methods for parameter estimates in the Rasch model. British Journal of Mathematical and Statistical Psychology 32:113-120
- Cole, N. S. 1973. Bias in selection. Journal of Educational Measurement 10:237-255
- Colonus, H. 1977. On Keats' generalization of the Rasch model. Psychometrika 42:443-445
- Conger, A. J. 1974. A revised definition for suppressor variables: A guide to their identification and interpretation. Educational and Psychological Measurement 34:35-46
- Coombs, C. H. 1953. On the use of objective examinations. Educational and Psychological Measurement 13:308-310
- Corder-Bolz, C. R. 1978. The evaluation of change: New evidence. Educational and Psychological Measurement 38:959-976
- Corno, L. 1979. A hierarchical analysis of selected naturally occurring aptitude-treatment interactions. American Educational Research Journal 16:391-409
- Cory, C. H. 1976. Relative utility of computerized versus paper-and-pencil tests for predicting job performance. Applied Psychological Measurement 1:551-564
- Cory, C. H., Rimland, B. & Bryson, R. A. 1977. Using computerized tests to measure new dimensions of abilities: An exploratory study. Applied Psychological Measurement 1:101-110
- Cronbach, L. J. 1976. Equity in selection--Where psychometrics and political philosophy meet. Journal of Educational Measurement 13:31-43
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. 1972. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley
- Cronbach, L. J., & Snow, R. E. 1977. Aptitudes and instructional methods. New York: Irvington Publishers
- Cross, L., & Frary, R. B. 1977. An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement 14:313-321
- Cross, L. H., & Frary, R. B. 1978. Empirical choice weighting under "guess"

- and "do not guess" instructions. Educational and Psychological Measurement 38:613-620
- Cudek, R., McCormick, D., & Cliff, N. 1979. Monte carlo evaluation of implied orders as a basis for tailored testing. Applied Psychological Measurement 3:65-74
- D'Agostino, R. B., & Cureton, E. E. 1975. The 27 percent rule revisited. Educational and Psychological Measurement 35:47-50
- Darlington, R. B. 1971. Another look at "cultural fairness." Journal of Educational Measurement 8:71-82
- Darlington, R. B. 1976. A defense of "rational" personnel selection and two new methods. Journal of Educational Measurement 13:43-52
- Darlington, R. B. 1978. Cultural test bias: Comment on Hunter and Schmidt. Psychological Bulletin 85:673-675
- Davison, M. L. 1977. On a metric, unidimensional unfolding model for attitudinal and developmental data. Psychometrika 42:523-548
- Davison, M. L. 1979. Testing a unidimensional, qualitative unfolding model for attitudinal or developmental data. Psychometrika 44:179-194
- Davison, M. L. 1980. A psychological scaling model for testing order hypotheses. British Journal of Mathematical and Statistical Psychology (in press)
- Davison, M. L., King, P. M., Kitchener, K. S., & Parker, C. A. 1980. The stage sequence concept in cognitive and social development. Developmental Psychology 16:121-131
- Davison, M. L., & Robbins, S. 1978. The reliability and validity of objective indices of moral development. Applied Psychological Measurement 2:391-404
- Davison, M. L., Robbins, S., & Swanson, D. B. 1978. Stage structure in objective moral judgments. Developmental Psychology 14:137-146
- Davison, M. L., & Thoma, S. J. 1980. CONSCAL: A FORTRAN program for testing structural hypotheses. Applied Psychological Measurement 4:8
- Dayton, C. M., & MacReady, G. B. 1976. A probabilistic model for validation of behavioral hierarchies. Psychometrika 41:189-204
- DeNisi, A. S., & Shaw, J. B. 1977. Investigation of the uses of self-reports of abilities. Journal of Applied Psychology 62:641-644
- Diamond, J. J. 1975. A preliminary study of the reliability and validity of a scoring procedure based upon confidence and partial information. Journal of Educational Measurement 12:129-133
- Dinero, T. E., & Haertl, E. 1977. Applicability of the Rasch model with vary-

- ing item discriminations. Applied Psychological Measurement 1:581-592
- Douglas, F. M., Khalil, A. K., & Farber, P. D. 1979. A comparison of classical and latent trait item analysis procedures. Educational and Psychological Measurement 39:337-352
- Douglas, G. A. 1978. Conditional maximum-likelihood estimation for a multiplicative binomial response model. British Journal of Mathematical and Statistical Psychology 31:73-83
- Downey, R. G. 1979. Item option weighting of achievement tests: Comparative study of methods. Applied Psychological Measurement 3:453-461
- Drösler, J. 1978. Extending the temporal range of psychometric prediction by optimal linear filtering of mental test scores. Psychometrika 43:533-550
- Dudek, F. J. 1979. The continuing misinterpretation of the standard error of measurement. Psychological Bulletin 86:335-337
- Duncan, G. T., & Milton, E. O. 1978. Multiple-answer multiple-choice test items: Responding and scoring through Bayes and minimax strategies. Psychometrika 43:43-57
- Dyck, W., & Plancke-Schuyten, G. 1976. Manipulations with multiple-choice tests: A probability function of a test score V. Educational and Psychological Measurement 36:259-262
- Eakin, R. R., & Long, C. A. 1977. Dodging the dilemma of true-false testing. Educational and Psychological Measurement 37:659-663
- Ebel, R. L. 1962. Content standard test scores. Educational and Psychological Measurement 22:15-25
- Ebel, R. L. 1975. Prediction? Validation? Construct validity? (Paper presented at the Content Validity II Conference, Bowling Green, Ohio)
- Ebel, R. L. 1978. The ineffectiveness of multiple true-false items. Educational and Psychological Measurement 38:37-44
- Echternacht, G. 1974. A quick method for determining test bias. Educational and Psychological Measurement 34:271-280
- Echternacht, G. 1975. The variances of empirically derived option score weights. Educational and Psychological Measurement 35:307-311
- Echternacht, G. 1976. Reliability and validity of item option weighting schemes. Educational and Psychological Measurement 36:301-309
- Einhorn, H. J., & Bass, A. R. 1971. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin 75:261-269
- Faggen-Steckler, J., McCarthy, K. A., & Tittle, C. K. 1974. A quantitative method for measuring sex "bias" in standardized tests. Journal of Educa-

tional Measurement 11:151-161

- Farrel, B. M. 1979. Task performance self-evaluations: An alternative selection procedure to traditional experience and training ratings. St. Paul, MN: Minnesota Department of Personnel Selection Research Unit
- Feldt, L. S. 1975. Estimation of the reliability of a test divided into two parts of unequal length. Psychometrika 40:557-561
- Flaughter, R. L. 1978. The many definitions of test bias. American Psychologist 33:671-679
- Flaughter, R. L., & Schrader, W. B. 1978. Eliminating differentially difficult items as an approach to test bias. Princeton, NJ: Educational Testing Service (ETS-RB-78-4)
- Fleiss, J. L. 1976. Comment on Overall and Woodward's asserted paradox concerning the measurement of change. Psychological Bulletin 83:774-775
- Forsyth, R. A. 1976. Estimating means via multiple matrix sampling: A note on the effect of selected data base characteristics. Educational and Psychological Measurement 36:275-282
- Forsyth, R. A. 1978a. A note on "Planning an experiment in the company of measurement error" by Levin and Subkoviak. Applied Psychological Measurement 2:379-383
- Forsyth, R. A. 1978b. Some additional comments on "Planning an experiment in the company of measurement error." Applied Psychological Measurement 2:386-387
- Gael, S., Grant, D. L., & Ritchie, R. J. 1975a. Employment test validation for minority and nonminority telephone operators. Journal of Applied Psychology 60:411-419
- Gael, S., Grant, D. L., & Ritchie, R. J. 1975b. Employment test validation for minority and nonminority clerks with work sample criteria. Journal of Applied Psychology 60:420-426
- Gibbons, J. D., Olkin, I., & Sobel, M. 1979. A subset selection technique for scoring items on a multiple-choice test. Psychometrika 34:259-270
- Glass, G. V. 1978. Standards and criteria. Journal of Educational Measurement 15:237-261
- Golding, S. L. 1977. Method variance, inadequate constructs or things that go bump in the night. Multivariate Behavioral Research 12:89-98
- Golding, S. L., & Seidman, E. 1974. Analysis of multitrait-multimethod matrices: A two step procedure. Multivariate Behavioral Research 9:479-496
- Goldman, R. D., & Hewitt, B. N. 1975. An investigation of test bias for Mexican-American college students. Journal of Educational Measurement 12:187-196

- Goldman, R. D., & Hewitt, B. N. 1976. Predicting the success of Black, Chicano, Oriental, and White college students. Journal of Educational Measurement 13:107-118
- Goldman, R. D., & Richards, R. 1974. The SAT prediction of grades for Mexican-American versus Anglo-American students at the University of California, Riverside. Journal of Educational Measurement 11:129-140
- Goldman, R. D., & Widawski, M. H. 1976. An analysis of types of errors in the selection of minority college students. Journal of Educational Measurement 13:185-200
- Green, B. F. 1978. In defense of measurement. American Psychologist 33:664-670
- Green, S. B., Lissitz, R. W., & Mulaik, S. A. 1977. Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement 37:827-838
- Greener, J. M., & Osburn, H. G. 1979. An empirical study of the accuracy of corrections for restriction in range due to explicit selection. Applied Psychological Measurement 3:31-41
- Grier, J. B. 1975. The number of alternatives for optimum test reliability. Journal of Educational Measurement 12:109-113
- Gross, A. L., Steckler-Faggen, J., & McCarthy, K. 1974. Statistical procedures for evaluating the practical utility of a moderator approach to prediction. Journal of Applied Psychology 59:578-582
- Gross, A. L., Faggen, J., & McCarthy, K. 1974. The differential predictability of the college performance of males and females. Educational and Psychological Measurement 34:363-365
- Gross, A. L., & Su, W. 1975. Defining a "fair" or "unbiased" selection mode: A question of utilities. Journal of Applied Psychology 60:345-351
- Guilford, J. P. 1967. The nature of human intelligence. New York: McGraw-Hill
- Guion, R. M. 1974. Open a new window: Validities and values in psychological measurement. American Psychologist 29:287-296
- Guion, R. M. 1977. Content validity--The source of my discontent. Applied Psychological Measurement 1:1-10
- Guion, R. M. 1978. Scoring of content domain samples: The problem of fairness. Journal of Applied Psychology 63:499-506
- Gullickson, A., & Hopkins, K. 1976. Interval estimation of correlation coefficients corrected for restriction of range. Educational and Psychological Measurement 36:9-25

- Gustafsson, J. E. 1979. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. Journal of Educational Measurement 16:153-158
- Guttman, L. 1944. A basis for scaling quantitative data. American Sociological Review 9: 139-150
- Hakstian, A. R., & Kansup, K. 1975. A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. Testing procedures. Journal of Educational Measurement 12:231-239
- Hakstian, A. R., & Whalen, T. E. 1976. A K-sample test for independent alpha coefficients. Psychometrika 41:219-231
- Haladyna, T. M. 1974. Effects of different samples on items and test characteristics of criterion-referenced tests. Journal of Educational Measurement 19:93-100
- Hambleton, R. K. 1978. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement 15:277-290
- Hambleton, R. K., & Cook, L. L. 1977. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement 14:75-95
- Hambleton, R. K., & Cook, L. L. 1980. The robustness of latent trait models and effect of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory
- Hambleton, R. K., Swaminathan, H., Algina, J., & Coulsen, D. B. 1978. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research 48:1-48
- Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. 1978. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research 48:467-510
- Hammer, W. C., Kim, S. J., Baird, L., & Bigoness, W. J. 1974. Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. Journal of Applied Psychology 59:705-711
- Hanna, G. S. 1975. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. Journal of Educational Measurement 12:175-178
- Harris, C. W., & Pearlman, A. P. 1978. An index for a domain of completion or short-answer items. Journal of Educational Statistics 3:285-303
- Healy, J. D. 1979. On Kristoff's test for a linear relation between true scores of two measures. Psychometrika 44:235-238

- Hilton, T. L. 1976. Intellectual status and intellectual growth again. Princeton, NJ: Educational Testing Service (ETS RB-76-29)
- Hively, W., Patterson, H. L., & Page, S. H. 1968. A "universe defined" system of arithmetic achievement testing. Journal of Educational Measurement 5:275-290
- Hoepfner, R. 1974. Published tests and needs of educational accountability. Educational and Psychological Measurement 34:103-109
- Hoffman, R. J. 1975. The concept of efficiency in item analysis. Educational and Psychological Measurement 35:621-640
- Hogan, R., DeSoto, C. B., & Solano, C. 1977. Traits, tests, and personality research. American Psychologist 32:255-264
- Hoogstraten, J. P. 1979. Pretesting as determinant of attitude change in evaluation research. Applied Psychological Measurement 3:25-30
- Hornke, L. L., & Sauter, M. B. 1980. A validity study of an adaptive test of reading comprehension. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory
- Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. 1979. Internal validity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. Applied Psychological Measurement 3:1-23
- Hsu, L. M. 1978. Determination of significance levels for tests of item validity. Educational and Psychological Measurement 38:209-211
- Hsu, L. M. 1979a. A comparison of three methods of scoring true-false tests. Educational and Psychological Measurement 39:785-790
- Hsu, L. M. 1979b. Ordering power of separate versus group true-false tests: Interaction of type of test with knowledge levels of examinees. Applied Psychological Measurement 3:529-536
- Hubert, L. J., & Baker, F. B. 1978. Analyzing the multitrait-multimethod matrix. Multivariate Behavioral Research 13:163-180
- Hubert, L. J., & Baker, F. B. 1979. A note on analyzing the multitrait-multimethod matrix: An application of a generalized proximity function comparison. British Journal of Mathematical and Statistical Psychology 32:179-184
- Huck, S. W. 1978a. A modification of Hoyt's analysis of variance reliability estimation procedure. Educational and Psychological Measurement 38:725-736
- Huck, S. W. 1978b. Handling "tied items" when using Lu's method of reliability estimation. Educational and Psychological Measurement 38:61-68

- Hunt, D. E. 1975. Person-environment interaction: A challenge found wanting before it was tried. Review of Educational Research 45:209-230
- Hunter, J. E., & Schmidt, F. L. 1974. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. American Psychologist 29:1-8
- Hunter, J. E., & Schmidt, F. L. 1976. Critical analysis of the statistical and ethical implications of various definitions of test bias. Psychological Bulletin 83:1053-1071
- Hunter, J. E., & Schmidt, F. L. 1978a. Bias in defining test bias: Reply to Darlington. Psychological Bulletin 85:675-676
- Hunter, J. E., & Schmidt, F. L. 1978b. Differential and single-group validity of employment tests by race. Journal of Applied Psychology 63:1-11
- Hunter, J. E., Schmidt, F. L., & Hunter, R. 1979. Differential validity of employment tests. Psychological Bulletin 86:721-735
- Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. 1977. Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology 62:245-261
- Huynh, H. 1976a. On consistency of decisions in criterion-referenced testing. Journal of Educational Measurement 13:253-265
- Huynh, H. 1976b. Statistical considerations of mastery scores. Psychometrika 41:65-78
- Huynh, H. 1979. Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics 4:231-246
- Huynh, H., & Perney, J. 1979. Determination of mastery scores when instructional units are linearly related. Educational and Psychological Measurement 39:317-323
- Ironson, G. H., & Subkoviak, M. 1979. A comparison of several methods of assessing test bias. Journal of Educational Measurement 16:209-226
- Jackson, D. N. 1975. Multimethod factor analysis: A reformulation. Multivariate Behavioral Research 10:259-276
- Jackson, D. N. 1977. Distinguishing trait and method variance in multitrait-multimethod matrices: A reply to Golding. Multivariate Behavioral Research 12:99-110
- Jackson, G. G. 1975. Comment on "Educational uses of tests with disadvantaged students." American Psychologist 30:88-92
- Jackson, P. H. 1979. A note on the relation between coefficient alpha and Guttman's "split-half" lower bounds. Psychometrika 44:251-252

- Jackson, P. H., & Agunwamba, C. C. 1977. Lower bounds for the reliability of the total score on a test composed of non-homogeneous items. I: Algebraic lower bounds. Psychometrika 42:567-578
- Jensema, C. 1976. A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement 36:705-715
- Jensema, C. J. 1977. Bayesian tailored testing and the influence of item bank characteristics. Applied Psychological Measurement 1:111-120
- Jensen, A. R. 1974. The effect of race of examiner on the mental test scores of white and black pupils. Journal of Educational Measurement 11:1-34
- Jensen, A. R. 1980. Bias in mental testing. New York: The Free Press
- Jepsen, D. A., & Grove, W. M. 1980. Stage order and dominance in adolescent vocational decision-making processes: An empirical test of the Tiedeman-O'Hara paradigm. (Manuscript submitted for publication, The University of Iowa)
- Joe, G. W., & Woodward, J. A. 1975. An approximate confidence interval for maximum coefficient alpha. Multivariate Behavioral Research 10:93-98
- Joe, G. W., & Woodward, J. A. 1976. Some developments in multivariate generalizability theory. Psychometrika 41:205-217
- Kaiser, H. F., & Michael, W. B. 1975. Domain validity and generalizability. Educational and Psychological Measurement 35:31-35
- Kaiser, H. F., & Michael, W. B. 1977. Little Jiffy factor scores and domain validities. Educational and Psychological Measurement 37:363-365
- Kalleberg, A. L., & Kleugel, J. R. 1975. Analysis of the multitrait-multimethod matrix: Some limitations and an alternative. Journal of Applied Psychology 60:1-9
- Kane, M., & Moloney, J. 1978. The effect of guessing on item reliability under answer-until-correct scoring. Applied Psychological Measurement 2:41-49
- Kansup, W., & Hakstian, A. R. 1975. A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. Journal of Educational Measurement 12:219-230
- Katzell, R. A., & Dyer, F. J. 1977. Differential validity revisited. Journal of Applied Psychology 62:137-145
- Katzell, R. A., & Dyer, F. J. 1978. On differential validity and bias. Journal of Applied Psychology 63:19-21
- Kearns, J., & Meredith, W. 1975. Methods for evaluating Bayes point estimates of latent trait scores. Psychometrika 40:373-394
- Keats, J. A. 1974. Applications of projective transformations of test theory.

Psychometrika 40:373-340

- Kingsbury, G. G., & Weiss, D. J. 1979a. An adaptive testing strategy for mastery decisions. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (Research Report 79-5)
- Kingsbury, G. G., & Weiss, D. J. 1979b. Relationships among achievement level estimates from three item characteristic curve scoring methods. Minneapolis: University of Minnesota, Department of Psychology. Psychometric Methods Program (Research Report 79-3)
- Kleinke, D. J. 1979. Systematic errors in approximations to the standard error of measurement. Applied Psychological Measurement 3:161-164
- Krus, D. J. 1977. Order analysis: An inferential model of dimensional analysis and scaling. Educational and Psychological Measurement 37:587-601
- Krus, D. J. 1978. Logical basis of dimensionality. Applied Psychological Measurement 2:323-331
- Krus, D. J., & Ceurvorst, R. W. 1979. Dominance, information, and hierarchical scaling of variance space. Applied Psychological Measurement 3:515-527
- Krus, D. J., & Weiss, D. J. 1976. Empirical comparison of factor and order analysis on prestructured and random data. Multivariate Behavioral Research 11:95-104
- Kwansa, K. B. 1974. Content validity and reliability of domain referenced tests. African Journal of Educational Research 1:73-79
- Levin, H. M. 1978. Educational performance standards: Image or substance. Journal of Educational Measurement 15:309-319
- Levin, J. R. 1974. A rotational procedure for separation of trait, method, and interaction factors in multitrait-multimethod matrices. Multivariate Behavioral Research 9:231-239
- Levin, J. R., & Subkoviak, M. J. 1977. Planning an experiment in the company of measurement error. Applied Psychological Measurement 1:331-338
- Levin, J. R., & Subkoviak, M. J. 1978. Correcting "Planning an experiment in the company of measurement error." Applied Psychological Measurement 2:382-385
- Levine, E. L., Flory, A. III, & Ash, R. A. 1977. Self-assessment in personnel selection. Journal of Applied Psychology 62:428-435
- Levine, M. V. 1976. The academic achievement test: Its historical context and social functions. American Psychologist 31:228-238
- Levine, M. V., & Drasgow, F. 1980. Appropriateness measurement: Basic principles and validating studies. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Min-

- nesota, Department of Psychology, Computerized Adaptive Testing Laboratory
- Levine, M. V., & Rubin, D. B. 1979. Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics 4:269-290.
- Lewis, C., Wang, M., & Novick, M. R. 1975. Marginal distributions for the estimation of proportions in m groups. Psychometrika 40:63-75
- Linn, R. L. 1973. Fair test use in selection. Review of Educational Research 43:139-161
- Linn, R. L. 1976a. In search of fair selection procedures. Journal of Educational Measurement 13:53-58
- Linn, R. L. 1976b. Response models and examinee behavior: A note on the lack of correspondence. Educational and Psychological Measurement 36:835-841
- Linn, R. L. 1978a. Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement 15:301-307
- Linn, R. L. 1978b. Single-group validity, differential validity, and differential prediction. Journal of Applied Psychology 63:507-512
- Linn, R. L., & Slinde, J. A. 1977. The determination of the significance of change between pre- and post-testing periods. Review of Educational Research 47:121-150
- Lissitz, R. W., & Schoenfeldt, L. F. 1974. Moderator subgroups for the estimation of educational performance. American Educational Research Journal 11:63-75
- Livingston, S. A. 1972. Criterion-referenced applications of classical test theory. Journal of Educational Measurement 9:13-26
- Livingston, S. A., & Wingersky, M. S. 1979. Assessing the reliability of tests used to make pass/fail decisions. Journal of Educational Measurement 16:247-260
- Loevinger, J. 1957. Objective tests as instruments of psychological theory. Psychological Reports 3:635-694 (Monograph Supplement No. 9)
- Lomax, R. G., & Algina, J. 1979. Comparison of two procedures for analyzing multitrait-multimethod matrices. Journal of Educational Measurement 16:177-186
- Lord, F. M. 1958. Some relations between Guttman's principal components of scale analyses and other psychometric theory. Psychometrika 23:291-296
- Lord, F. M. 1975a. Formula-scoring and number-right scoring. Journal of Educational Measurement 12:7-11
- Lord, F. M. 1975b. Relative efficiency of number-right and formula scores. British Journal of Mathematical and Statistical Psychology 28:46-50

- Lord, F. M. 1975c. The 'ability' scale in item characteristic curve theory. Psychometrika 14:205-217
- Lord, F. M. 1977a. A broad-range test of verbal ability. Applied Psychological Measurement 1:95-100
- Lord, F. M. 1977b. Optimal number of choices per item--A comparison of four approaches. Journal of Educational Measurement 14:33-38
- Lord, F. M. 1977c. Practical applications of item characteristic curve theory. Journal of Educational Measurement 14:117-138
- Lord, F. M. 1977d. Some item analysis and test theory for a system of computer-assisted test construction. Applied Psychological Measurement 1:447-455
- Lord, F. M., & Stocking, M. L. 1976. An interval estimate for making statistical inferences about true scores. Psychometrika 41:79-87
- Lovett, H. T. 1977. Criterion referenced reliability estimated by ANOVA. Educational and Psychological Measurement 37:21-29
- Lumsden, J. 1976. Test theory. Annual Review of Psychology 27:257-280
- Lumsden, J. 1977. Person reliability. Applied Psychological Measurement 1:477-482
- Lumsden, J. 1978. Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology 31:19-26
- Lumsden, J. 1980. Variations on a theme by Thurstone. Applied Psychological Measurement 4:1-7
- Maccoby, E. E., & Jacklin, C. N. 1974. The psychology of sex differences. Stanford: Stanford University Press
- McBride, J. R. 1977. Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement 1:121-140
- McDonald, R. P. 1978. Generalizability in factorable domains: "Domain validity and generalizability." Educational and Psychological Measurement 38:75-79
- McFatter, R. M. 1979. The use of structural equation models in interpreting regression equations including suppressor and enhancer variables. Applied Psychological Measurement 3:123-135
- McNemar, Q. 1975. On so-called test bias. American Psychologist 30:848-851
- McQuitty, L. L. 1976. Comprehensive analysis of test item variance. Educational and Psychological Measurement 36:51-84
- Macready, G. B., & Dayton, C. M. 1977. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics 2:99-120

- Mai-Dalton, R. R., Feldman-Summers, S., & Mitchell, T. R. 1979. Effect of employee gender and behavioral style on evaluations of male and female bank executives. Journal of Applied Psychology 64:221-226
- Marco, G. L. 1977. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement 14:139-160
- Marco, G. L., Petersen, N. S., & Stewart, E. E. 1980. A test of the adequacy of curvilinear score equating models. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory
- Marshall, J. L., & Haertel, E. H. 1976. The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. (Unpublished manuscript, University of Wisconsin)
- Medley, D. M., & Quirk, T. J. 1974. The application of a factorial design to the study of cultural bias in general culture items on the national teacher examination. Journal of Educational Measurement 11:235-246
- Mehrens, W. A., & Lehmann, I. J. 1978. Measurement and evaluation in education and psychology (2nd ed.). Chicago: Holt, Rinehart, & Winston
- Mellenbergh, G. J., Kelderman, H., Stijlen, J. G., & Zondag, E. 1979. Linear models for the analysis and construction of instruments in a facet design. Psychological Bulletin 86:766-776
- Mellenbergh, G. J., & van der Linden, W. J. 1979. The internal and external optimality of decisions based on tests. Applied Psychological Measurement 3:257-273
- Mercer, J., & Lewis, J. F. 1978. SOMPA: System of multicultural pluralistic assessment (Ages 5-11). New York: The Psychological Corporation
- Meskauskas, J. A. 1976. Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. Review of Educational Research 46:133-158
- Messick, S. 1975. The standard problem: Meaning and values in measurement and evaluation. American Psychologist 30:955-966
- Millman, J., & Popham, W. J. 1974. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement 11:137-138
- Mischel, W. 1977. On the future of personality measurement. American Psychologist 32:246-265
- Molenaar, W. 1977. On Bayesian formula scores for random guessing in multiple-choice tests. British Journal of Mathematical and Statistical Psychology 30:79-89

- Morrison, D. G., & Brockway, G. 1979. A modified beta-binomial model with applications to multiple-choice and taste tests. Psychometrika 44:427-442
- Moss, J. D., & Brown, F. G. 1979. Sex bias and academic performance: An empirical study. Journal of Educational Measurement 16:197-202
- Myerberg, N. J. 1979. The effect of item stratification on the estimation of the mean and variance of universe scores in multiple matrix sampling. Educational and Psychological Measurement 39:57-68
- Myers, C. T. 1975. Test fairness: A comment on fairness in statistical analyses. Princeton, NJ: Educational Testing Service (ETS RB-75-12)
- Neill, J. A., & Jackson, D. N. 1976. Minimum redundancy item analysis. Educational and Psychological Measurement 36:123-134
- Nevo, B. 1977. Using item test-retest stability (ITRS) as a criterion for item selection: An empirical study. Educational and Psychological Measurement 31:847-852
- Nicewander, W. A. 1975. A relationship between Harris factors and Guttman's sixth lower bound to reliability. Psychometrika 40:197-203
- Nicewander, W. A., & Price, J. M. 1978. Dependent variable reliability and the power of significance tests. Psychological Bulletin 85:405-409
- Nicewander, W. A., Price, J. M., Mendoza, J. L., & Henderson, D. 1977. The attenuation paradox and the distribution of ability. British Journal of Mathematical and Statistical Psychology 30:204-209
- Norris, L. 1976. The SIGI prediction system: Predicting college grades with and without tests. Princeton, NJ: Educational Testing Service (ETS RB-76-26)
- Norris, L., & Chapman, W. 1976. A test-free approach to prediction for guidance. Princeton, NJ: Educational Testing Service (ETS RB-76-32)
- Novick, M. R., & Ellis, D. D., Jr. 1977. Equal opportunity in educational and employment selection. American Psychologist 32:306-320
- Novick, M. R., & Jackson, P. H. 1974. Further cross-validation analysis of the Bayesian M-group regression method. American Educational Research Journal 11:77-85
- Oakland, T. (Ed.). 1977. Psychological and educational assessment of minority children. New York: Bruner/Mazel
- O'Connor, E. J., Wexley, K. N., & Alexander, R. A. 1975. Single-group validity: Fact or fallacy. Journal of Applied Psychology 60:352-355
- Oosterhof, A. C. 1977. Similarity of various item discrimination indices. Journal of Educational Measurement 13:145-150

- Osburn, H. G. 1968. Item sampling for achievement testing. Educational and Psychological Measurement 28:95-104
- Osburn, H. G., & Greener, J. M. 1978. Optimal sampling strategies for validation studies. Journal of Applied Psychology 63:602-608
- Overall, J. E., & Woodward, J. A. 1975. Unreliability of difference scores. Psychological Bulletin 82:85-86
- Overall, J. E., & Woodward, J. A. 1976. Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. Psychological Bulletin 83:776-777
- Pandey, T. J., & Shoemaker, D. M. 1975. Estimating moments of universe scores and associated standard errors in multiple matrix sampling for all item-scoring procedures. Educational and Psychological Measurement 35:567-581
- Pandey, T. N., & Hubert, L. 1975. An empirical comparison of several interval estimation procedures for coefficient alpha. Psychometrika 40:169-181
- Panell, R. C., & Laabs, G. J. 1979. Construction of a criterion-referenced diagnostic test for an individualized instruction program. Journal of Applied Psychology 64:255-261
- Pearlman, K. 1980. Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin 87:1-28
- Perline, R., Wright, B. D., & Wainer, H. 1979. The Rasch model as additive conjoint measurement. Applied Psychological Measurement 3:237-255
- Petersen, N. S., & Novick, M. R. 1976. An evaluation of some models for culture-fair selection. Journal of Educational Measurement 13:3-29
- Pfeifer, C. M., Jr. 1976. Relationship between scholastic aptitude, perception of university climate, and college success for black and white students. Journal of Applied Psychology 61:341-347
- Pine, S. M., & Weiss, D. J. 1977. Effects of item characteristics on test fairness. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (Research Report 76-5)
- Pinney, G. W. 1979. Eighth graders below average in reading. Minneapolis Tribune, November 11, p. 1B.
- Pohlmann, J. T., & Beggs, D. L. 1974. A study of the validity of self-reported measures of academic growth. Journal of Educational Measurement 11:115-119
- Poizner, S. B., Nicewander, W. A., & Gettys, C. F. 1978. Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. Applied Psychological Measurement 2:83-96

- Popham, W. J. 1975. Educational evaluation. Englewood Cliffs, NJ: Prentice-Hall
- Popham, W. J. 1978. As always provocative. Journal of Educational Measurement 15:297-300
- Pugh, R. C., & Brunza, J. J. 1975. Effects of a confidence-weighted scoring system on measures of test reliability and validity. Educational and Psychological Measurement 35:73-78
- Raffeld, P. 1975. The effects of Guttman weights on the reliability and predictive validity of objective tests when omissions are not differentially weighted. Journal of Educational Measurement 12:179-185
- Raju, N. S. 1977a. On estimating test variance in multiple matrix sampling. Educational and Psychological Measurement 37:621-625
- Raju, N. S. 1977b. A generalization of coefficient alpha. Psychometrika 42:549-565
- Raju, N. S. 1979. Note on two generalizations of coefficient alpha. Psychometrika 44:347-349
- Ray, M. L., & Heeler, R. M. 1975. Analysis techniques for exploratory use of the multitrait-multimethod matrix. Educational and Psychological Measurement 35:255-265
- Reckase, M. D. 1979. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics 4:207-230
- Ree, M. J. 1979. Estimating item characteristic curves. Applied Psychological Measurement 3:371-385
- Ree, M. J., & Jensen, H. E. 1980. The effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory
- Reeb, M. 1976. Differential test validity for ethnic groups in the Israeli Army and the effects of educational level. Journal of Applied Psychology 61:253-261
- Reed, C. W. 1976. Statistical issues raised by Title IX requirements on admissions procedures. Princeton, NJ: Educational Testing Service (ETS RB-76-25)
- Reid, F. 1977. An alternative scoring formula for multiple-choice and true-false tests. Journal of Educational Research 60:335-339
- Reilly, R. R. 1975. Empirical option weighting with a correction for guessing. Educational and Psychological Measurement 35:613-619

- Reilly, R. R., Zedeck, S., & Tenopyr, M. L. 1979. Validity and fairness of physical ability tests for predicting performance in craft jobs. Journal of Applied Psychology 64:262-274
- Rentz, F. R., & Bashaw, W. L. 1977. The National Reference Scale for Reading: An application of the Rasch model. Journal of Educational Measurement 14:161-179
- Richards, J. M., Jr. 1975. A simulation study of the use of change measures to compare educational programs. American Educational Research Journal 12:299-311
- Roark, A. C. 1978. Post-Bakke admission plan adopted at Davis. Chronicle of Higher Education, October 30, p. 13
- Rock, D. A., Werts, C. E., Linn, R. L., & Jöreskog, K. G. 1977. A maximum likelihood solution to the errors in variables and errors in equations model. Multivariate Behavioral Research 12:187-197
- Roe, R. A. 1979. The correction for restriction of range and the difference between intended and actual selection. Educational and Psychological Measurement 39:551-559
- Rovinelli, R. J., & Hambleton, R. K. 1977. On the use of test specialists in the assessment of criterion-referenced test item validity. Dutch Journal of Educational Research 2:49-60
- Rowley, G. L., & Traub, R. F. 1977. Formula scoring, number-right scoring, and test-taking strategy. Journal of Educational Measurement 14:15-22
- Rozeboom, W. W. 1978. Domain validity--Why care? Educational and Psychological Measurement 38:81-88
- Rubin, D. B., & Thayer, D. 1978. Relating tests given to different samples. Psychometrika 43:3-10
- Rybak, R. T. 1980. State's affirmative action program is confused. Minneapolis Tribune, March 3, pp. 1A, 4A
- Samejima, F. 1977a. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. Psychometrika 42:163-191
- Samejima, F. 1977b. A use of the information function in tailored testing. Applied Psychological Measurement 1:233-247
- Samejima, F. 1977c. Effects of individual optimization in setting the boundaries of dichotomous items on accuracy of estimation. Applied Psychological Measurement 1:77-94
- Samejima, F. 1977d. Weakly parallel tests in latent trait theory with some criticisms of classical test theory. Psychometrika 42:193-198
- Samuda, R. J. 1975. Psychological testing of American minorities: Issues and

- consequences. New York: Harper & Row
- Sands, W. A., Alf, E. F., & Abrahams, N. M. 1978. Correction of validity coefficients for direct restriction in range occasioned by univariate selection. Journal of Applied Psychology 63:747-750
- Sawyer, R. L., Cole, N. S., & Cole, J. R. 1976. Utilities and the issue of fairness in a decision theoretic model. Journal of Educational Measurement 13:59-76
- Scheuneman, J. 1979. A method of assessing bias in test items. Journal of Educational Measurement 16:143-152
- Schmidt, F. L. 1977. The Urry method of approximating the item parameters of latent trait theory. Educational and Psychological Measurement 37:613-620
- Schmidt, F. L., & Hunter, J. E. 1977. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology 62:529-541
- Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. 1979. Impact of valid selection procedures on work force productivity. Journal of Applied Psychology 64:609-626
- Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. 1979. Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology 32:257-281
- Schmitt, N. 1978. Path analysis of multitrait-multimethod matrices. Applied Psychological Measurement 2:157-173
- Schmitt, N., Coyle, B. C., & Saari, B. B. 1977. A review and critique of analyses of multitrait-multimethod matrices. Multivariate Behavioral Research 12:447-478
- Schmitt, N., Melton, P. M., & Bylenga, C. 1978. Sex differences in validity for academic and employment criteria and different types of predictors. Journal of Applied Psychology 63:145-150
- Schoenfeldt, L. F., Schoenfeldt, B. B., Acker, S. R., & Perlson, M. R. 1976. Content validity revisited: The development of a content-oriented test of industrial reading. Journal of Applied Psychology 61:581-588
- Schulman, R. S. 1976. Correlation and prediction in ordinal test theory. Psychometrika 41:329-340
- Schulman, R. S. 1978. Individual distributions under ordinal measurement. Psychometrika 43:19-29
- Schulman, R. S., & Haden, R. L. 1975. A test theory for ordinal measurements. Psychometrika 40:455-472
- Scriven, M. How to anchor standards. Journal of Educational Measurement 15:273-275

- Sedere, M. V., & Feldt, L. S. 1977. The sampling distributions of the Kristof reliability coefficient, the Feldt coefficient, and Guttman's Lambda-2. Journal of Educational Measurement 14:53-62
- Serlin, R. C., & Kaiser, H. F. 1978. A method for increasing the reliability of a short multiple-choice test. Educational and Psychological Measurement 38:337-340
- Sewall, G., Carey, J., Simons, P. E., & Lord, F. M. 1980. Tests: How good? How fair? Newsweek, February 18, pp. 97-104
- Shoemaker, D. M. 1974. Toward a framework for achievement testing. Review of Educational Research 44:127-147
- Silverman, B. J., Barton, F., & Lyon, M. 1976. Minority group status and bias in college and admissions criteria. Educational and Psychological Measurement 36:401-407
- Simas, K., & McCarrey, M. 1979. Impact of recruiter authoritarianism and applicant sex on evaluation and selection decisions in a recruitment interview analogue study. Journal of Applied Psychology 64:483-491
- Slatker, M. J., Crehan, K. D., & Koehler, R. A. 1975. Longitudinal studies of risk-taking on objective examinations. Educational and Psychological Measurement 35:97-105
- Slinde, J. A., & Linn, R. L. 1977a. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement 15:23-35
- Slinde, J. A., & Linn, R. L. 1977b. Vertically equated tests: Fact or phantom? Journal of Educational Measurement 14:23-32
- Slinde, J. A., & Linn, R. L. 1979a. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement 16:159-165
- Slinde, J. A., & Linn, R. L. 1979b. The Rasch model, objective measurement, equating and robustness. Applied Psychological Measurement 3:437-452
- Smeiser, C. B., & Ferguson, R. L. 1978. Performance of black and white students on test materials containing content based on black and white cultures. Journal of Educational Measurement 15:193-200
- Spinetti, J. P., & Hambleton, R. K. 1977. A computer simulation study of tailored testing strategies for objectives-based instructional programs. Educational and Psychological Measurement 37:139-158
- Strasler, G. M., & Raeth, P. G. 1977. An internal consistency estimate for criterion-referenced tests. (Paper presented at the annual meeting of the National Council on Measurement in Education)

- Subkoviak, M. J. 1976. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement 13:265-275
- Subkoviak, M. J. 1978. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement 15:111-116
- Subkoviak, M. J., & Levin, J. R. 1977. Fallibility of measurement and the power of a statistical test. Journal of Educational Measurement 14:47-52
- Swaminathan, H., Hambleton, R. K., & Algina, J. 1974. Reliability of criterion-referenced tests. Journal of Educational Measurement 11:263-267
- Swaminathan, H., Hambleton, R. K., & Algina, J. 1975. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement 12:87-98
- ten Berge, J. M. F., & Zegers, F. E. 1978. A series of lower bounds to the reliability of a test. Psychometrika 43:575-579
- Terwilliger, J. S. 1977. Assigning grades--Philosophical issues and practical recommendations. Journal of Research and Development in Education 10:21-39
- Terwilliger, J. S., & Lele, K. 1977. Some relationships among internal consistency, reproducibility and homogeneity. Journal of Educational Measurement 16:101-108
- Tesser, A., & Krauss, H. 1976. On validating a relationship between constructs. Educational and Psychological Measurement 36:111-121
- Thissen, D. M. 1976. Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement 13:201-214
- Thorndike, R. L. 1971. Concepts of culture-fairness. Journal of Educational Measurement 8:63-70
- Tinsley, H. E. A., & Dawis, R. V. 1975. An investigation of the Rasch simple logistic model: Sample free item and test calibration. Educational and Psychological Measurement 35:325-339
- Tinsley, H. E. A., & Dawis, R. V. 1977. Test-free person measurement with the Rasch simple logistic model. Applied Psychological Measurement 1:483-487
- Tobias, S. 1976. Achievement treatment interactions. Review of Educational Research 46:61-74
- Trabin, T. E., & Weiss D. J. 1979. The person response curve: Fit of individuals to item characteristic curve models. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (Research Report 79-7)
- Traub, R. E., & Fisher, C. W. 1977. On the equivalence of constructed-response and multiple-choice tests. Applied Psychological Measurement 1:355-369

- Tryon, R. C. 1957. Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin 54:229-249
- Tucker, L. R. 1966. Some mathematical notes on three-mode factor analysis. Psychometrika 31:279-311
- U.S. Equal Employment Opportunity Commission. 1970. Guidelines on employment selection procedures. Washington, DC: Author
- Urry, V. W. 1976. Ancillary estimators for the item parameters of mental test models. In W. A. Gorham (Chair), Computers and testing: Steps toward the inevitable conquest. Washington DC: U.S. Civil Service Commission, Personnel Research and Development Center (PS-76-1)
- Urry, V. W. 1977. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement 14:181-196
- Vale, C. D., & Weiss, D. J. 1975. A comparison of information functions of multiple-choice and free-response vocabulary items. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (Research Report 77-2)
- van der Linden, W. J. 1979. Binomial test models and item difficulty. Applied Psychological Measurement 3:401-411
- van der Linden, W. J., & Mellenbergh, G. J. 1978. Coefficients of tests from a decision theoretic point of view. Applied Psychological Measurement 2:119-134
- Velicer, W. F. 1978. Suppressor variables and the semi-partial correlation coefficient. Educational and Psychological Measurement 38:953-958
- Wade, T. C., & Baker, T. B. 1977. Opinions and use of psychological tests: A survey of clinical psychologists. American Psychologist 32:874-882
- Wainer, H., Fairbank, D., & Hough, R. L. 1978. Predicting the impact of simple and compound life change events. Applied Psychological Measurement 2:315-324
- Wainer, H., & Wright, B. D. 1980. Robust estimation in the Rasch model. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory
- Waller, M. 1980. Estimating ability within the two-parameter logistic latent trait model in the presence of a non-symmetric distribution of ability. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory
- Warren, J. R. 1976. Prediction of college achievement among Mexican-American students in California. Princeton, NJ: Educational Testing Service (ETS RB-76-22)

- Waters, B. K. 1977. An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement 1:141-152
- Weiss, D. J. 1974. Strategies of adaptive ability measurement. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (Research Report 74-5)
- Weiss, D. J. (Ed.) Final Report: Computerized ability testing, 1972-1975. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program
- Weiss, D. J. (Ed.) 1978. Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program
- Weiss, D. J. (Ed.) 1980. Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory
- Wen, S-S. 1975. The relationship between verbal-meaning test scores and degree of confidence in item responses. Journal of Educational Measurement 12:197-200
- Werts, C. E., & Hilton, T. L. 1977. Intellectual status and intellectual growth. American Educational Research Journal 14:137-146
- Werts, C. E., Linn, R. L., & Jöreskog, K. G. 1977. A simplex model for analyzing academic growth. Educational and Psychological Measurement 37:745-756
- Werts, C. E., Rock, R. D., Linn, R. L., & Jöreskog, K. G. 1978. A general method of estimating the reliability of a composite. Educational and Psychological Measurement 933-938
- White P. O. 1976. A note on Keats' generalization of the Rasch model. Psychometrika 41:405-407
- Whitely, S. E. 1977. Models, meanings and misunderstandings: Some issues in applying Rasch's theory. Journal of Educational Measurement 14:227-235
- Whitely, S. E. 1979. Estimating measurement error on highly speeded tests. Applied Psychological Measurement 3:141-154
- Whitely, S. E., & Dawis, R. V. 1975. A model for psychometrically distinguishing aptitude from ability. Educational and Psychological Measurement 35:51-66
- Whitely, S. E., & Dawis, R. V. 1976. The influence of test context on item difficulty. Educational and Psychological Measurement 36:329-337
- Wilcox, R. R. 1976. A note on the length and passing scores of a mastery test. Journal of Educational Statistics 1:359-364

- Wilcox, R. R. 1977. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics 2:289-307
- Wilcox, R. R. 1978a. Estimating true score in the compound binomial error model. Psychometrika 43:245-262
- Wilcox, R. R. 1978b. A note on decision theoretic coefficients for tests. Applied Psychological Measurement 2:609-613
- Wilcox, R. R. 1979a. Applying ranking and selection techniques to determine the length of a mastery test. Educational and Psychological Measurement 39:13-27
- Wilcox, R. R. 1979b. Achievement tests and latent structure models. British Journal of Mathematical and Statistical Psychology 32:61-71
- Wilcox, R. R. 1979c. An alternative interpretation of three stability models. Educational and Psychological Measurement 39:311-315
- Wilcox, R. R. 1979d. A lower bound to the probability of choosing the optimal passing score for a mastery test when there is an external criterion. Psychometrika 44:245-249
- Wilcox, R. R. 1979e. Estimating the parameters of the beta-binomial distribution. Educational and Psychological Measurement 39:527-535
- Wilcox, R. R. 1979f. On false-positive and false-negative decisions with a mastery test. Journal of Educational Statistics 4:59-73
- Williams, R. H., & Zimmerman, D. W. 1977. The reliability of difference scores when errors are correlated. Educational and Psychological Measurement 37:679-689
- Williams, R. L. 1975. Black intelligence test of cultural homogeneity: Manual of directions. St. Louis, MO: Author
- Wilson, K. M. 1978. Predicting the long-term performance in college of minority and nonminority students: A comparative analysis in two collegiate settings. Princeton, NJ: Educational Testing Service (ETS RB-78-6)
- Winne, P. H. 1977. Aptitude treatment interactions in an experiment on teacher effectiveness. American Educational Research Journal 14:389-409
- Wood, R. L. 1976. Inhibiting blind guessing: The effect of instructions. Journal of Educational Measurement 13:297-307
- Wood, R. L. 1978. Fitting the Rasch model--A heady tale. British Journal of Mathematical and Statistical Psychology 31:27-32
- Wood, R. L., Wingersky, M. S., & Lord, F. M. 1976. LOGIST: A computer program for estimating examinee ability and item characteristic parameters. Princeton, NJ: Educational Testing Service (ETS RB-76-6)

- Woodhouse, B., & Jackson, P. H. 1977. Lower bounds for the reliability of the total score on a test composed of non-homogeneous items. II: A search procedure to locate the greatest lower bound. Psychometrika 42:579-591
- Woodson, M. I. C. E. 1974. The issue of item and test variance for criterion-referenced tests: A reply. Journal of Educational Measurement 11:139-140
- Woodward, J. A., & Bentler, P. M. 1978. A statistical lower bound to population reliability. Psychological Bulletin 85:1323-1326
- Wright, B. D. 1977a. Misunderstanding the Rasch model. Journal of Educational Measurement 14:219-225
- Wright, B. D. 1977b. Solving measurement problems with the Rasch model. Journal of Educational Measurement 14:97-116
- Wright, B. D., & Douglas, G. A. 1977a. Best procedures for sample-free item analysis. Applied Psychological Measurement 1:281-295
- Wright, B. D., & Douglas, G. A. 1977b. Conditional versus unconditional procedures for sample-free item analysis. Educational and Psychological Measurement 37:573-586
- Wright, B. D., & Stone, M. A. 1980. Best test design. Chicago: Mesa
- Wright, R. J., & Bean, A. G. 1974. The influence of socioeconomic status on the predictability of college performance. Journal of Educational Measurement 11:277-284
- Zimmerman, D. W. 1976. Test theory with minimal assumptions. Educational and Psychological Measurement 36:85-96

